

Article

An Interpretable Vision-Language Framework for Evaluating the Uncanny Valley Effect of XR Humanoid Characters

Xiner Li ¹, Yi Xiao ^{1,*}, Jinhao Qiao ², Yan Zheng ² and Chi-Sing Leung ³¹ School of Design, Hunan University, Changsha 410082, China; xinerli@hnu.edu.cn² School of Artificial Intelligence and Robotics, Hunan University, Changsha 410082, China; qiaojh@hnu.edu.cn (J.Q.); yanzheng@hnu.edu.cn (Y.Z.)³ Department of Electronic Engineering, City University of Hong Kong, Hong Kong 999077, China; eeleungc@cityu.edu.hk

* Correspondence: yixiao_csee@hnu.edu.cn

Abstract

As AI-generated humanoid characters are increasingly used in virtual, augmented, and mixed reality applications, evaluating the Uncanny Valley Effect (UVE) is crucial for immersive user experience. Existing evaluation methods map visual features to affective scores, offering limited interpretability regarding which visual cues are associated with affinity judgments. Among the theoretical perspectives proposed to explain the UVE, perceptual conflict provides a visual-cue-oriented perspective for analyzing whether local-feature realism supports a coherent overall human-likeness impression and how this is reflected in affinity judgments, yet this perspective is rarely incorporated into interpretable UVE assessment. Thus, we propose UVE-Perception Chain-of-Thought (UVE-PCoT), a vision-language framework for interpretable UVE evaluation from a perceptual-conflict-oriented perspective. UVE-PCoT organizes assessment through a structured perceptual decomposition, including assessments of overall human-likeness, local-feature realism, perceptual conflict, and affinity. To provide supervision, we construct UVE-R, a structured rationale dataset with image-grounded, rating-consistent rationales linking visual cue observations, cue-level inconsistency analysis, and affinity judgments. Results show that UVE-PCoT improves affinity prediction and cue-level explanation over general-purpose multimodal large language models and ablations. Our approach operationalizes this perceptual-conflict-oriented perspective into an interpretable framework, advancing UVE evaluation from black-box scoring to explanatory analysis and providing cue-level insights for XR character assessment and revision.

Keywords: uncanny valley effect; XR humanoid character; vision-language model; chain-of-thought reasoning; interpretable



Academic Editor: Andrea Asperti

Received: 21 April 2026

Revised: 23 June 2026

Accepted: 3 July 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid development of virtual reality (VR), augmented reality (AR), and mixed reality (MR), humanoid characters are increasingly becoming central interaction agents in immersive environments, with expanding applications in virtual companionship [1], digital guidance [2], education [3], and intelligent assistance [4]. Meanwhile, recent advances in generative artificial intelligence, particularly diffusion-based image and avatar generation, have substantially improved the efficiency of creating highly realistic digital humans, avatars, and interactive characters [5–7]. However, greater realism does not necessarily lead to better user experience. When a humanoid character appears highly human-like

yet still contains subtle imperfections, it may evoke eeriness, discomfort, or rejection, a phenomenon commonly referred to as the Uncanny Valley Effect (UVE) [8,9]. As illustrated in Figure 1, user affinity does not increase linearly with human-likeness. Once a character reaches a critical level of realism and becomes nearly human but still perceptually imperfect, user acceptance may decline sharply [10–13]. Although the UVE is inherently multimodal and can be influenced by motion, sound, and expression [14,15], static facial appearance remains an important factor in early visual evaluation for XR humanoid characters because it often forms the user's first impression before extended interaction begins. For character designers, the practical challenge is therefore not only to measure users' affective responses, but also to identify which specific visual cues are associated with affinity judgments, so that these cues may inform subsequent design analysis and revision.

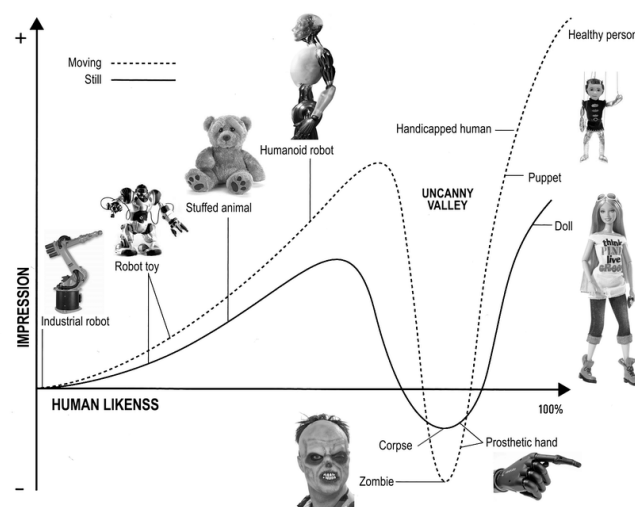


Figure 1. The hypothesized Uncanny Valley Effect.

Existing UVE assessment methods, however, face a fundamental trade-off between interpretability and practical efficiency. Participant-dependent approaches can identify visual features associated with UVE-related responses by directly collecting subjective judgments or physiological responses. These methods mainly include subjective assessment using standardized questionnaires, such as the Godspeed Scales [16], and objective measurement using physiological signals, including fMRI [17], EEG [18], and eye-tracking [19]. Because they are grounded in human responses, such approaches can provide useful insights into which features may be associated with affinity judgments. However, they are inherently post hoc and resource intensive, since each new humanoid character design often requires additional user studies or experimental measurement.

By contrast, computational approaches aim to improve efficiency through mathematical modeling or artificial neural networks (ANNs). Mathematical models typically describe the UVE as a polynomial relationship between human-likeness and affective response, making it possible to estimate the human-likeness range in which the valley emerges [20,21]. ANN-based methods further support automatic prediction from visual input. For example, CLIP (Contrastive Language-Image Pre-Training)-based networks have shown promise in aligning machine perception with human perception in the UVE domain [22]. Nevertheless, mathematical models can only characterize the overall location of the valley, while ANN-based methods largely operate as black boxes and offer limited insight into which specific visual cues are associated with the final prediction. As a result, existing methods still struggle to provide both efficient evaluation and interpretable cue-level insights for humanoid character design analysis.

Recent advances in Vision-Language Models (VLMs), especially when combined with Chain-of-Thought (CoT) reasoning, have created new opportunities for the interpretable assessment of subjective perceptual phenomena such as the UVE. Unlike conventional ANNs that directly map visual input to a final score, CoT allows a model to express its prediction through intermediate explanatory steps. This makes the model output more transparent and provides a possible path from black-box prediction to interpretable analysis. Such a capability is particularly valuable for UVE evaluation, where designers need not only a final affective score, but also an explanation of which cues may have contributed to that judgment. Related studies have shown the promise of this idea in subjective evaluation tasks. For example, the ArtCoT framework proposed by Jiang and Chen shows that decomposing aesthetic judgment into intermediate steps, such as content analysis and style analysis, helps improve the alignment between model outputs and human aesthetic judgments [23]. Ke et al. further show that vision-language joint modeling can leverage natural language information to compensate for the semantic limitations of learning from scores alone in subjective aesthetic tasks [24]. These findings suggest that VLMs with CoT-based explanations provide a useful basis for interpretable UVE evaluation.

In practice, however, directly applying free-form CoT in general VLMs to UVE evaluation often produces errors. To better understand why these errors occur, it is necessary to clarify whether they mainly stem from insufficient visual perception capabilities, a lack of UVE-oriented inference annotations in the training data, or the absence of an explicit interpretive framework for relating visual cues to user affinity judgments. Perceptual conflict offers a visual-cue-oriented perspective for interpreting how visual cues in humanoid characters are reflected in UVE-related responses and affinity judgments. In the context of static humanoid faces, this perspective emphasizes that user evaluation does not depend only on isolated visual features, but also on whether local-feature realism, such as the eyes, skin texture, facial proportions, and hair, jointly supports a coherent overall human-likeness impression [25–27]. When different local cues convey inconsistent levels of realism, users may perceive the character as less coherent, which may be associated with lower affinity judgments. To illustrate this issue, we present representative examples showing how general VLMs can describe visible appearance cues but still produce unreliable UVE evaluations.

Firstly, to examine whether these errors can be attributed solely to insufficient visual perception, we evaluate general VLMs on the basic perceptual dimensions involved in UVE evaluation from a perceptual-conflict-oriented perspective, including overall human-likeness and the realism of local facial features such as geometry, texture, and hair. These dimensions correspond to the two perceptual levels required for judging whether local cues jointly support a coherent overall human-like impression. As shown in Section 4.2.1 (Table 3), several models achieve reasonable, though not perfect, performance on these perceptual dimensions, often exhibiting high alignment with human ratings. This suggests that general VLMs can capture many individual visual cues relevant to UVE evaluation. However, when the task shifts from assessing these perceptual attributes to final UVE-related affinity evaluation, performance drops substantially, with R^2 values often becoming negative (Section 5.2, Table 5). This contrast indicates that the errors cannot be explained solely by insufficient visual perception capabilities.

Secondly, these errors may also be related to the lack of UVE-specific supervision in general VLMs. To examine this possibility, we further consider the image-only fine-tuning setting, in which models are trained using humanoid images and their human-rated affinity labels, but without intermediate perceptual rationales or cue-level inconsistency rationales. In other words, this setting provides image-affinity supervision only. As shown in Section 5.2 (Table 5), image-only fine-tuning improves affinity prediction compared with

free-form general VLMs. For example, Qwen3-VL-4B achieves an MAE of 0.4904 and an R^2 of 0.5450, while InternVL3.5-4B achieves an MAE of 0.4687 and an R^2 of 0.5796. These results indicate that UVE-specific affinity labels provide useful task supervision. However, the remaining errors are still substantial, suggesting that image-affinity supervision alone is not sufficient to fully address the difficulty of UVE evaluation.

To further illustrate the remaining limitation, we selected two representative VLMs, InternVL3.5 [28] and Qwen3.5-Plus [29], compared their predicted affinity scores and generated explanations with human ratings for humanoid characters. Figure 2 summarizes these comparisons, highlighting both the potential and the limitations of current VLMs with free-form CoT in UVE evaluation. In Example 1, despite the explicit mechanical attributes of the target character, human raters assigned low affinity scores due to the intense stare of the blue eyes, the frozen expression, and an overall impression of lifelessness. Conversely, both models relied more heavily on the definition of the UVE, assuming that a clearly robotic character resides outside the typical category of being almost human. Furthermore, they interpreted features such as oversized eyes as cuteness cues, thereby underestimating the potential discomfort evoked in humans. Example 2 clarifies this limitation further. When presented with a highly human-like character possessing subtle abnormalities, human raters gave low affinity ratings and referred to unease associated with the synthetic smile, waxy skin texture, lifeless eyes, and unnatural hair. In contrast, while both models identified certain abnormal cues, they did not adequately explain how these local inconsistencies were reflected in lower affinity ratings. These two cases suggest that VLMs with free-form CoT can detect visible anomalies, but still struggle to represent how multiple cues jointly contribute to human affinity judgments. Rather than evaluating whether local-feature realism cues jointly support a coherent overall human-like impression, they tend to treat these cues as isolated visual attributes, making it difficult to explain how visual cues are reflected in affinity judgments.

Example 1  Question: Based on the uncanny valley effect hypothesis, please give a score for the character's affinity in this image (1 = extremely low affinity, 2 = low affinity, 3 = neutral, 4 = high affinity, 5 = very high affinity) and explain your reasoning.	Answer from InternVL3.5 Affinity Rating: 3 Reasoning: This character blends human-like features (eyes, brows, mouth) with overtly non-human elements—a smooth white shell, forehead sensor, and simplified mechanical mouth. While this partial anthropomorphism edges toward the uncanny valley, large blue eyes and rounded contours tap into notoriety principles, conveying harmlessness and boosting affinity. However, hollow-looking eyes (lacking iris detail) dampen emotional resonance, while static brows and a rigid mechanical mouth fail to convey dynamic expression. These lifeless qualities offset the friendly elements, limiting the score to the neutral range.	Answer from Qwen3.5-Plus Affinity Rating: 4 Reasoning: This character design effectively avoids the uncanny valley by embracing a distinctly non-human aesthetic. Its robotic white head, exaggerated blue eyes, and smooth, textureless surface clearly signal it as a fictional creation, preventing any uncomfortable "almost human" perception. The high affinity is driven by a strong "baby schema"—the oversized eyes and rounded head instinctively trigger a nurturing response. By committing to a stylized, cartoon-like appearance rather than realism, the design ensures viewers engage with it purely on terms of charm and artistic appeal.	Answer from Human 1 Affinity Rating: 1 Reasoning: These huge blue eyes are terrifying. They stare right through you with a creepy, fake human intensity. With the frozen eyebrows and heavy eyelids, the face looks totally vacant. The way the mouth hangs open makes the jaw look broken. It's deeply unsettling.	Answer from Human 2 Affinity Rating: 2 Reasoning: The character's blue eyes and eyebrows convey a sense of melancholy, almost sorrowful, while its mouth resembles that of a puppet—stiff and lifeless. The overall effect feels subtly unsettling.
Example 2  Question: Based on the uncanny valley effect hypothesis, please give a score for the character's affinity in this image (1 = extremely low affinity, 2 = low affinity, 3 = neutral, 4 = high affinity, 5 = very high affinity) and explain your reasoning.	Answer from InternVL3.5 Affinity Rating: 3 Reasoning: The character's facial structure, including proportions and smile, closely resembles a real human, yet overly smooth skin, stiff hair, and lifeless eyes introduce subtle uncanny valley effects that slightly reduce its overall appeal. However, the friendly expression—conveyed through a gentle smile and soft facial lines—counters this sense of alienation, preventing an excessively low rating. Striking a balance between realism and stylization, the design retains simplified features like clear contours and relatively flat shading, positioning it in a neutral territory that avoids the deepest pitfalls of hyper-realism while maintaining approachability.	Answer from Qwen3.5-Plus Affinity Rating: 2 Reasoning: Based on the uncanny valley hypothesis, this human-like character receives a likability score of 2 out of 5. The character closely mimics human appearance but falls into the uncanny valley due to subtle flaws—such as unnaturally smooth skin, a stiff smile, and slight proportion inconsistencies. These imperfections create an "almost real but not quite" effect that triggers viewer discomfort.	Answer from Human 1 Affinity Rating: 2 Reasoning: This character's smile looks really fake, almost like the mouth was Photoshopped on. It's just so stiff and unnatural, not like a real person's smile at all. And his skin has a plastic or waxy texture, nothing like real skin. The hair looks fake too, like it's just one solid piece stuck on his head, like a bad wig. The whole thing just makes me feel really uncomfortable.	Answer from Human 2 Affinity Rating: 1 Reasoning: This character feels like a human wearing a mask. It has been forced into a smile. His eyes are completely lifeless and there's no hint of a genuine smile in them. That stretched-out mouth doesn't suggest he's actually happy either. It's frightening. His skin and hair also look fake. The whole thing looks like a mannequin, and it's terrifying.

Figure 2. Comparison of affinity ratings and explanations from InternVL3.5, Qwen3.5-Plus, and human raters.

To further understand why this gap arises, we then examined the generated intermediate explanations for Example 1 in Figure 2, as shown in Figure 3. InternVL3.5 begins by defining the task and then analyzing local visual features, including the smooth plastic-like texture, oversized but hollow-looking blue eyes, simplified static facial structure, and a forehead sensor. It subsequently applies UVE theory, maps these cues onto a predefined scoring framework, and concludes that the character lies in a neutral but slightly odd range. Because it recognizes both neotenic friendliness and lifeless mechanical stiffness, it assigns

a final score of 3. Qwen3.5-Plus also begins by defining the task, but then explicitly recalls the general UVE framework before analyzing the same character as a rounded, cartoon-like humanoid robot with large blue eyes and clear mechanical signals. It interprets these features mainly as stylized and baby-schema-like cues, judges that the character falls into a safe robot/cartoon zone, and therefore assigns a higher affinity score of 4. Overall, both models rely on generic UVE theory recall, local-feature matching, and score mapping. They tend to aggregate isolated cue-level impressions, such as mechanical stiffness, stylized cuteness, or lifelessness, into a final score without explicitly examining whether these local-feature realism cues jointly support a coherent overall human-like impression. As a result, their explanations remain plausible at the descriptive level, but they are not organized by an explicit UVE-oriented interpretive structure for relating local visual cues to overall human-likeness impressions and affinity judgments.

The reasoning steps of InternVL3.5	The reasoning steps of Qwen3.5-Plus
● Step 1: Define the task Identify the user's request: rate the affinity of the humanoid robot in the image on a scale from 1 to 5 based on the Uncanny Valley Effect, and provide corresponding justifications ● Step 2: Analyze visual features Identify that the object is a humanoid robot head with a smooth white plastic/ceramic texture (lacking real skin and iris detail). It has very large bright blue eyes but hollow-looking pupils, simplified eyebrows, a static rigid mouth structure, and a prominent rectangular sensor on the forehead. The design blends human-like features (eyes, brows, mouth) with overtly non-human elements—smooth shell, mechanical mouth, forehead sensor. ● Step 3: Apply the Uncanny Valley Effect theory Map these features onto the UVE framework. This partial anthropomorphism edges toward the uncanny valley: the character sits in a middle zone where human-like layout meets unnatural materials and lifeless expressions. Large eyes and rounded contours tap into neoteny principles, conveying harmlessness and boosting affinity. However, hollow-looking eyes (lacking iris detail) dampen emotional resonance, while static brows and a rigid mechanical mouth fail to convey dynamic expression. These lifeless qualities offset the friendly elements. ● Step 4: Determine the rating interval Establish a scoring reference: 1 = extremely eerie/terrifying; 2 = clearly uncomfortable; 3 = neutral but slightly odd; 4 = fairly likable/warm; 5 = fully human or highly appealing cartoon character. The character does not fall into the deepest valley nor is it truly likable. The mixture of neotenic cuteness and lifeless mechanical stiffness places it in the neutral range (around 3), rather than the lower discomfort zone (2) or the warm zone (4). ● Step 5: Formulate the rating conclusion Final rating: 3 (neutral but slightly odd). Core reasoning: The character's large eyes and rounded shape provide some harmless cuteness, but hollow eyes, static brows, and a rigid mechanical mouth collectively limit emotional resonance, keeping the score in the neutral range.	● Step 1: Define the task Identify the user's request: rate the affinity of the human-like character on a scale from 1 to 5 based on the Uncanny Valley Effect. ● Step 2: Recall the theoretical framework Actively invoke the UVE definition: when a robot or animated character reaches a certain level of human-likeness but falls short of being fully realistic, it tends to evoke unease. Conversely, if the entity is clearly a machine or cartoon, it generally maintains higher affinity. Low human-likeness is safe; medium may fall into the valley; high human-likeness can see affinity rise again. ● Step 3: Analyze visual features Identify that the character is a white humanoid robot head with large blue eyes, a square mechanical forehead component, and obvious mechanical structures. Its overall shape is rounded and cartoonish, giving it a toy-like or child-friendly robot appearance. The design embraces a distinctly non-human aesthetic—robotic white head, exaggerated eyes, smooth textureless surface—clearly signaling it as a fictional creation. ● Step 4: Apply the theory to determine the interval This design effectively avoids the uncanny valley because it never attempts realism. The exaggerated blue eyes and rounded head trigger a strong "baby schema," instinctively evoking a nurturing response. By committing to a stylized, cartoon-like appearance rather than realism, the character falls closer to the "clearly a robot/cartoon" zone, where affinity is naturally high. The mechanical forehead component adds slight oddness but does not push it into the valley. ● Step 5: Formulate the rating conclusion After comprehensive judgment, the model gives a moderately high affinity rating: 4 out of 5. Core reasoning: The character's design ensures viewers engage with it purely on terms of charm and artistic appeal. Its high affinity is driven by neoteny and a clear non-human aesthetic, avoiding any uncomfortable "almost human" perception. Local mechanical features only slightly weaken the score, keeping it below a perfect 5 but well above neutral.

Figure 3. Comparison of the generated intermediate explanations between InternVL3.5 and Qwen3.5-Plus on UVE evaluation.

To support interpretable UVE evaluation, especially the identification of visual cues associated with users' affinity judgments, we draw on this perceptual-conflict-oriented perspective and propose UVE-PCoT, a structured perceptual decomposition framework. Rather than treating affinity prediction as a direct mapping from appearance to score, UVE-PCoT organizes UVE assessment into four stages. First, Overall Human-likeness Assessment establishes a reference along the human-likeness continuum. Second, Local-Feature Realism Assessment analyzes facial geometry, texture, and hair to extract local perceptual cues. Third, Perceptual Conflict Detection assesses whether local-feature realism jointly supports a coherent overall human-like impression or reveals inconsistencies across cues. Finally, Affinity Evaluation integrates the preceding stages to produce the affinity prediction. To provide structured supervision, we construct UVE-R, a structured rationale annotation dataset with image-grounded rationales generated using human rating scores as anchors during annotation construction. These rationales provide rating-consistent and image-grounded cue-level explanatory annotations that link perceptual observations, cue-level inconsistency analysis, and affinity judgments in humanoid character evaluation. Accordingly, UVE-R should be understood as providing structured rationale supervision for model training, rather than as reconstructing the internal psychological reasoning of human raters. Experimental results show that UVE-PCoT improves affinity prediction and cue-level explanation compared with general VLMs and ablated variants. By making

these perceptual relationships explicit, UVE-PCoT can provide cue-level insights that may support subsequent assessment and revision of humanoid character designs.

2. Related Work

This section reviews prior work from three perspectives that motivate our framework: psychological and perceptual accounts of the UVE, computational methods for its measurement and assessment, and recent advances in Vision-Language Models (VLMs) with general Chain-of-Thought (CoT) reasoning.

2.1. Psychological and Perceptual Mechanisms of the Uncanny Valley Effect

Early explanations of the UVE were largely grounded in evolutionary accounts, which interpret uncanny responses as adaptive reactions to biologically relevant threats. The pathogen avoidance hypothesis suggests that subtle abnormalities in appearance resemble cues of disease or genetic defects, thereby eliciting aversive behavioral immune responses [30,31]. Similarly, the mortality salience hypothesis proposes that highly lifelike yet non-living figures evoke implicit reminders of death, which in turn generate discomfort [32]. While these accounts help explain why uncanny stimuli may be affectively aversive, they provide limited insight into how such responses are formed through perceptual and cognitive processing.

Subsequent research has increasingly shifted toward cognitive and perceptual explanations, emphasizing that the UVE cannot be fully explained as a direct consequence of human-likeness alone. Instead, negative responses are more likely to emerge when human-like cues fail to form a coherent and stable perceptual whole. One influential line of work highlights expectation violation: a highly human-like appearance induces stronger expectations about how a character should move or behave, and mismatched behavioral or dynamic cues generate cognitive conflict [17]. Closely related accounts focus on categorization difficulty, arguing that eeriness arises when a character lies near the human and non-human boundary and resists stable category assignment [33]. Other studies further refine this view by showing that the UVE may be driven more specifically by perceptual inconsistency across features, such as realistic textures combined with stylized geometry or other mismatched realism cues [26,34,35]. At a higher social-cognitive level, the mind perception hypothesis suggests that highly human-like characters invite attributions of agency or mentality that become ontologically incompatible with their mechanical nature [36].

Taken together, these accounts suggest that human-likeness alone is insufficient to explain UVE-related affinity judgments. Humanoid characters with similar levels of overall human-likeness may still receive very different affinity evaluations, indicating that the UVE should also be interpreted in terms of how local visual cues support or disrupt a coherent overall human-like impression. From this perspective, overall human-likeness serves as a high-level perceptual reference for evaluating the consistency of local-feature realism. When local cues such as facial geometry, skin texture, eyes, or hair convey inconsistent levels of realism and fail to jointly support this overall impression, this inconsistency can be understood as perceptual conflict. Such conflict may make the character appear less coherent and may be associated with lower affinity judgments. This visual-cue-oriented account of perceptual conflict provides the theoretical motivation for our interpretable UVE evaluation framework, which examines whether local visual cues jointly support or disrupt the overall human-like impression before producing the final affinity judgment.

2.2. Measurement and Computational Assessment of the Uncanny Valley Effect

2.2.1. Participant-Dependent Measurement

Traditional UVE research has largely relied on human subject experiments, combining subjective reports with objective physiological measures to assess affective responses. Subjective approaches typically employ multi-dimensional rating scales, such as the Godspeed indices, to capture users' perceptions of humanoid characters across dimensions including human-likeness, eeriness, and likability [37]. These methods provide direct evidence of human judgments and remain important for testing theoretical claims. However, they mainly yield holistic outcome measures and offer limited insight into which specific visual cues are associated with affinity judgments or uncanny responses. Their results may also be influenced by cultural background, individual preference, and semantic ambiguity, which can reduce consistency across studies.

To capture more implicit aspects of UVE-related perception, researchers have also incorporated behavioral and neurophysiological measures. Eye-tracking studies show that observers tend to fixate longer on regions such as the eyes and mouth when viewing uncanny characters [38]. Functional magnetic resonance imaging (fMRI) studies report stronger activation in the parietal cortex and broader action observation network for uncanny stimuli, suggesting increased neural processing costs under violated expectations [17]. Electroencephalography (EEG) studies likewise identify event-related signals, such as the N400, associated with perceptual inconsistency and error detection [18]. While these methods provide valuable evidence about the cognitive processes underlying the UVE, they are costly, time consuming, and difficult to scale. As a result, participant-dependent approaches are not readily suited to rapid and repeated evaluation in contemporary digital character design, especially in workflows that require iterative generation and refinement.

2.2.2. Mathematical Formalization and Computational Modeling

To improve efficiency and formalize the structure of the UVE, computational approaches have attempted to model the relationship between human-likeness and affinity in quantitative terms. Early work often relied on polynomial regression to fit Mori's original valley-shaped curve. Although such studies are useful for testing whether a non-linear relationship exists, the resulting curve shapes vary considerably across datasets. Some studies report a clear uncanny valley pattern [38,39], whereas others support alternative forms such as an uncanny cliff or multiple local minima [40,41]. These inconsistencies reflect the strong dependence of polynomial fitting on stimulus composition and sampling density, which limits its value not only for general UVE assessment, but also for identifying which specific cues are associated with reduced affinity in practical design settings [42].

Beyond descriptive curve fitting, Moore provided a formal Bayesian account of the UVE [25]. In this view, observers do not respond to human-likeness as a single scalar property. Rather, they integrate multiple perceptual cues to infer whether a character should be categorized as human or non-human. When these cues support a consistent category judgment, perception remains relatively stable. However, when different cues provide conflicting evidence, perceptual uncertainty and tension may increase, which in turn can evoke negative affect. This account is theoretically important because it reframes the UVE as a problem of multi-cue integration and perceptual conflict, rather than as a simple response to increasing realism. However, MacDorman and Chattopadhyay showed that reducing the consistency of human realism across visual features increases eeriness and coldness, whereas increasing category uncertainty alone does not necessarily produce the UVE [26]. This suggests that UVE-related discomfort is not merely a matter of uncertainty about whether a character belongs to a human or non-human category, but is closely associated with inconsistencies among the realism levels of local features. Similarly, recent accounts have

further emphasized that the effect of feature deviations on UVE judgments depends on how these deviations interact with the overall human-like impression [27]. Taken together, these accounts suggest that UVE-related responses cannot be fully explained by human-likeness alone, but can also be interpreted from the perspective of whether local cues with different levels of realism jointly support a coherent overall human-like impression. Such inconsistency can be understood as perceptual conflict and may be associated with lower affinity judgments.

2.2.3. Machine Learning-Based Assessment

More recently, UVE assessment has increasingly moved toward data-driven methods. Early work often formulated the task as a binary classification problem, aiming to predict whether a humanoid character would elicit a comfortable or uncomfortable response from human observers. For example, Dal Molin et al. used Support Vector Machines (SVMs) with handcrafted descriptors, such as Hu moments and the Histogram of Oriented Gradients (HOG), to distinguish between comfortable and uncomfortable responses [43]. Ensemble methods have also been explored to improve robustness in sentiment classification [44]. These studies demonstrate the feasibility of automated UVE-related prediction, but their reliance on manually designed features limits both representational flexibility and generalization across diverse humanoid appearances.

To overcome the limitations of handcrafted descriptors, later studies turned to end-to-end models based on large pre-trained representations. In particular, Igaue and Hayashi investigated CLIP for UVE-related perception by constructing image continua between non-human objects and human faces and measuring their semantic associations with affective terms such as “eerie” and “weird” [22]. Their results suggest that large-scale vision-language representations can capture affective patterns related to the UVE and, to some extent, align with human judgments of eeriness. However, this ability largely reflects recognition of broad non-human versus human categories, rather than evaluating whether the realism of different features collectively supports a coherent human-like impression. Because CLIP maps visual content into a globally aligned embedding space, it is less suited to characterizing subtle cross-feature conflicts in facial geometry, texture appearance, and other fine-grained visual cues [26]. More broadly, existing machine learning approaches improve scalability, but most still lack cue-level interpretability and remain limited in identifying which visual cues are associated with human affinity judgments.

2.3. Vision-Language Models for Subjective Perceptual Evaluation

Recent advances in Vision-Language Models (VLMs) have created new opportunities for subjective visual evaluation. Compared with earlier visual models, VLMs combine visual representation learning with language-based reasoning, making them suitable for tasks that require not only perceptual recognition but also human-centered interpretation. Recent studies show that VLMs can engage with subjective image properties such as aesthetics, memorability, and emotional evocativeness, although their judgments still remain imperfectly aligned with human perception [45,46]. These findings suggest that VLMs provide a promising basis for modeling subjective responses, but alignment with human evaluative judgments remains a fundamental challenge.

Chain-of-Thought (CoT) reasoning provides one possible route toward improving such alignment. By decomposing complex judgments into intermediate steps, CoT can make model inference more transparent and improve reasoning quality [47]. This idea has also been extended to subjective visual tasks such as image aesthetic assessment, where multi-stage reasoning helps models analyze visual cues at different levels [48]. However, existing CoT-based frameworks are mostly designed for general reasoning or broad subjective evaluation. As a result, they may describe or evaluate individual visual cues, but they do

not explicitly organize these cues into a UVE-specific interpretive structure that relates local-feature realism to overall human-likeness impressions and affinity judgments. In particular, existing CoT-based approaches provide limited support for explaining whether multiple local cues jointly support a coherent overall human-like impression. This limitation motivates the need for a structured perceptual decomposition framework for interpretable affinity prediction and cue-level explanation.

3. Operationalizing Perceptual Conflict in UVE Evaluation

This section examines whether a rating-derived operational proxy of perceptual conflict shows statistical patterns consistent with the perceptual-conflict-oriented account of UVE evaluation. Because UVE-PCoT is informed by this perspective, this analysis provides empirical motivation for incorporating a perceptual-conflict-oriented structure into an interpretable evaluation framework, rather than serving as a procedure for constructing training labels or direct supervision signals. In this study, perceptual conflict is operationalized as a theory-guided proxy derived from human ratings, rather than as a directly observed psychological state. Following the perceptual-conflict-oriented account, we assume that affinity judgments are not determined by overall human-likeness alone, but may also be associated with whether local visual cues support a coherent overall human-likeness impression. To operationally examine this relationship, we first construct a residual-based perceptual conflict index from the inconsistency between overall human-likeness and local-feature realism. We then use hierarchical regression to examine whether this index provides additional explanatory power for affinity ratings beyond overall human-likeness alone. We also use grouped regression and visualization to examine whether the association between this index and affinity ratings varies across different levels of human-likeness. These analyses do not directly measure the full psychological process of perceptual conflict, nor is the resulting index used as a direct supervision signal for UVE-PCoT. Rather, the results provide empirical motivation for the subsequent design choice of incorporating overall human-likeness assessment, local-feature realism assessment, perceptual conflict detection, and affinity evaluation into UVE-PCoT.

3.1. Dataset

Our analysis is conducted on the humanoid character dataset introduced by Li et al. [42]. This dataset is particularly suitable for the present study because it provides ratings not only for overall human-likeness and affinity, but also for multiple dimensions of local realism. This annotation structure makes it possible to examine whether affinity judgments are associated not only with overall human-likeness, but also with the consistency between overall human-likeness and local-feature realism. The three local dimensions were selected to balance perceptual coverage and annotation feasibility. Geometry, texture, and hair capture major sources of realism in static face appearance, corresponding respectively to structural form, surface material appearance, and hair-related appendage cues. More fine-grained ratings of individual parts, such as eyes, nose, mouth, eyebrows, eyelashes, and skin details, could provide more detailed information, but would substantially increase annotator workload and fatigue and may reduce rating consistency. Therefore, these three dimensions provide a practical compromise between cue-level interpretability and reliable large-scale annotation.

The dataset contains 1000 static facial images spanning a broad spectrum of human-likeness, including robots, computer-animated characters, wax figures, real dolls, and real humans. Such diversity enables the dataset to cover a wide range of humanoid appearances. All annotations are given on a five-point scale. For example, an overall human-likeness score of 1 indicates a clear departure from human appearance, whereas a score of 5 indicates

an appearance that is nearly indistinguishable from a real human. The meanings of the annotated dimensions are summarized as follows:

- Overall Human-Likeness: This variable captures the overall impression of how human-like a character appears and is related to the horizontal axis in Mori's formulation of the UVE. The term "overall" is used to distinguish this holistic judgment from subsequent evaluations of individual local features.
- Local-Feature Realism: These scores reflect the perceived realism of specific facial components, indicating how closely they resemble those of a real human. Following a common decomposition in face perception and aesthetic evaluation, local realism is divided into geometry, texture, and hair-related cues:
 - Geometry Realism: The perceived realism of the structural aspects of the face, excluding skin texture and hair-related attributes. This includes overall facial proportions, the size and relative arrangement of facial features, and the anatomical plausibility of face shape.
 - Texture Realism: The perceived realism of skin appearance, including skin color, smoothness, wrinkles, pores, moles, and other surface-level details.
 - Hair Realism: The perceived realism of hair-related attributes, including head hair, eyebrows, eyelashes, beards, and other facial hair, as well as their color, texture, and material appearance.
- Affinity: The primary dependent variable in this study, reflecting the observer's subjective affective response to the character.

3.2. From Prior Theory to a Rating-Derived Proxy of Perceptual Conflict

Moore proposes that UVE-related responses cannot be explained by human-likeness alone, but may also involve perceptual tension when multiple cues provide conflicting support for category judgment [25]. In Moore's Bayesian account, this perceptual tension is not only described conceptually but is also formalized through a variance-based representation of cue disagreement. Specifically, when different cues provide divergent evidence for human or non-human category inference, the variance among cue-based estimates increases, indicating stronger cue disagreement. However, MacDorman and Chattopadhyay further show that reducing realism consistency across features increases eeriness and decreases warmth, whereas category uncertainty alone is not supported as the primary explanation [26]. Thus, these studies suggest that UVE-related responses can be interpreted not only in terms of human-likeness or category ambiguity, but also in terms of whether multiple cues form a coherent overall percept.

In addition, directly adopting Moore's original formulation is limited by a mismatch between its category-inference setting and the data structure used in this study. Although Moore's model provides a variance-based formalization of cue disagreement, its posterior probability difference cannot be directly implemented in the empirical setting of this study. This is because the original formulation requires prior distributions, likelihood functions, and posterior probabilities over discrete category judgments. In contrast, the dataset used in this study consists of continuous human ratings of overall human-likeness, local-feature realism, and affinity, rather than discrete human/non-human categorization data. Therefore, we retain Moore's core idea of using variance to capture cue disagreement, but adapt it to the continuous rating structure of our dataset. Because the perceptual-conflict-oriented account used in this study concerns whether local-feature realism cues jointly support a coherent overall human-like impression, the operational proxy needs to capture not only the absolute realism level of each cue, but also how differently these cues deviate from the overall human-likeness reference. Therefore, we first estimate the expected realism level of each local cue under a given overall human-likeness score and then calculate the

residual deviation of each cue from this expectation. The operational perceptual conflict index is then computed as the variance among these residual deviations. In this sense, the index is not an ad hoc metric, but a theory-guided residual variance proxy that adapts Moore's variance-based view of cue disagreement to the continuous rating structure of our dataset. Let S denote the overall human-likeness score and let R_i denote the realism of local cue i , where $i \in \text{geometry, texture, hair}$. For each cue, we estimate its expected realism level given the overall human-likeness of the character as Equation (1):

$$\hat{R}_i = \alpha_i S + \beta_i, \quad (1)$$

where $\hat{R}_i$ represents the level of local realism that cue i is expected to exhibit under the character's overall human-like impression. We then define the cue-specific deviation as Equation (2):

$$\epsilon_i = R_i - \hat{R}_i, \quad (2)$$

which quantifies the extent to which the observed realism of a local cue departs from the level implied by the character's overall human-likeness. Based on these deviations, perceptual conflict (PC) is operationalized as the dispersion of local-cue deviations:

$$PC = \text{Var}(\epsilon_{\text{geometry}}, \epsilon_{\text{texture}}, \epsilon_{\text{hair}}). \quad (3)$$

A larger PC value indicates that local cues deviate from the holistic reference in a more uneven and less coordinated manner, suggesting a lower degree of rating-level perceptual coherence. In this sense, our formulation is not a direct reproduction of Moore's Bayesian model or MacDorman and Chattopadhyay's controlled realism inconsistency manipulation. Rather, it is a theory-guided operational proxy adapted to the available rating structure, allowing us to examine whether rating-level cue inconsistency is associated with users' affinity ratings in the human rating dataset, without claiming to measure the full psychological process of perceptual conflict.

Next, we examine whether this operational perceptual conflict index shows statistical patterns consistent with the perceptual-conflict-oriented account in the current dataset. Specifically, we examine two questions:

- First, does the operationalized perceptual conflict index provide additional explanatory power for affinity ratings beyond the global effect of overall human-likeness?
- Second, does the association between this index and affinity ratings vary across different levels of human-likeness?

To address these questions, we use hierarchical regression to test the incremental explanatory value of the operational perceptual conflict index for affinity ratings, and grouped regression with visualization of the human-likeness and affinity relationship to examine how this association changes across the human-likeness continuum. These analyses provide empirical motivation for organizing the subsequent UVE-PCoT perceptual decomposition around a perceptual-conflict-oriented structure, rather than serving as direct supervision for the model or as proof of its internal reasoning process. We acknowledge that this operational proxy is simplified and may not fully capture complex nonlinear interactions among features or individual differences in feature weighting. However, the purpose of this analysis is not to exhaustively model the full psychological process of perceptual conflict. Instead, it aims to examine whether a rating-derived proxy of cue inconsistency is meaningfully associated with affinity judgments within the current dataset, thereby providing empirical motivation for constructing the structured rationale design in

UVE-PCoT around overall human-likeness assessment, local-feature realism assessment, perceptual conflict detection, and affinity evaluation.

3.3. Results

We first examined whether the operational perceptual conflict index is associated with affinity ratings beyond the global effect of overall human-likeness. To this end, we conducted a hierarchical regression analysis with two models. Model 1 followed Li et al. [42] and adopted the cubic polynomial of overall human-likeness (S , S^2 , and S^3) as the baseline, since that study identified the cubic function as the best-fitting model for the human-likeness and affinity relationship in this dataset. Model 2 extended this baseline by further adding the operationalized perceptual conflict index. As shown in Table 1, Model 1 accounted for a substantial proportion of the variance in affinity ratings ($R^2 = 0.617$, $F(3, 996) = 534.482$, $p < 0.001$). After adding the operational PC index, the explanatory power increased to $R^2 = 0.698$ ($F(4, 995) = 576.012$, $p < 0.001$), with a significant incremental $\Delta R^2 = 0.081$ ($F_{\text{change}} = 269.058$, $p < 0.001$). The PC term showed a significant negative coefficient ($\beta = -0.297$, $t = -16.403$, $p < 0.001$), indicating that stronger deviations of local realism cues from the holistic human-likeness reference were associated with lower affinity ratings beyond the effect of overall human-likeness alone.

Table 1. Hierarchical regression analysis for predicting affinity ratings.

Variable	Model 1 (Baseline)				Model 2 (with Perceptual Conflict)			
	<i>B</i>	<i>SE</i>	β	<i>t</i>	<i>B</i>	<i>SE</i>	β	<i>t</i>
Constant	3.786	0.283	—	13.378 ***	4.346	0.254	—	17.144 ***
<i>S</i>	−2.975	0.359	−3.406	−8.297 ***	−3.639	0.321	−4.166	−11.339 ***
S^2	1.116	0.134	7.442	8.326 ***	1.376	0.120	9.171	11.458 ***
S^3	−0.102	0.015	−3.375	−6.701 ***	−0.130	0.014	−4.296	−9.533 ***
<i>PC</i>	—	—	—	—	−0.099	0.006	−0.297	−16.403 ***
R^2	0.617				0.698			
Adj. R^2	0.616				0.697			
ΔR^2	—				0.081 ***			
<i>F</i> Change	534.482 ***				269.058 ***			

Note: *** $p < 0.001$. *B*—unstandardized coefficient; *SE*—standard error; β —standardized coefficient; *t*—*t* statistic; R^2 —coefficient of determination; Adj. R^2 —adjusted coefficient of determination; ΔR^2 —increment in explained variance relative to Model 1; *F* Change—change in model fit associated with the addition of *PC*. VIF for *PC* in Model 2 is 1.20.

Overall, this result suggests that the rating-derived PC index captures variance in affinity ratings that is not accounted for by overall human-likeness alone. At the dataset level, this pattern is consistent with the perceptual-conflict-oriented account that uneven deviations in local-feature realism may be associated with weaker support for a coherent overall human-like impression and with lower affinity ratings. Therefore, the PC index is used in this section as a theory-guided operational proxy to examine whether the human rating data show statistical patterns consistent with this perspective. However, this index should not be interpreted as a direct measurement of the full psychological process of perceptual conflict. It is a rating-derived residual variance measure computed from overall human-likeness and local-feature realism ratings, rather than an independently collected user rating of perceived conflict. For the same reason, it is not used as a direct supervision signal for UVE-PCoT. Its quantitative information is derived from existing rating dimensions and does not provide independent image-level supervision about which visual cues are perceived as inconsistent or how such cue-level inconsistency is reflected in affinity evaluation. Accordingly, the PC index provides empirical motivation within

the present dataset, while the following section introduces structured, image-grounded rationale supervision that allows UVE-PCoT to organize the relationships among overall human-likeness, local-feature realism, cue-level inconsistency analysis, and final affinity evaluation in a data-driven manner.

We next examined whether the association between the operational perceptual conflict index and affinity ratings varied across different levels of overall human-likeness. Grouped regression analyses were conducted for three ranges of overall human-likeness: Low ($S < 2.5$), Mid ($2.5 \leq S < 3.5$), and High ($S \geq 3.5$). As shown in Table 2, the effect of perceptual conflict changed systematically across the continuum. In the low-humanness regime, perceptual conflict was positively associated with affinity ratings ($\beta = 0.230, t = 8.519, p < 0.001$). In contrast, the coefficient became significantly negative in the mid-humanness regime ($\beta = -0.167, t = -9.278, p < 0.001$) and remained negative in the high-humanness regime ($\beta = -0.114, t = -11.400, p < 0.001$). This pattern indicates that the effect of perceptual conflict on affinity varies with the overall level of human-likeness. When human-likeness is low, local inconsistencies may be tolerated or interpreted as stylization. However, as a character becomes more human-like, similar inconsistencies are more likely to disrupt the coherent human-like impression and reduce affinity.

Table 2. Grouped regression results: association between operational perceptual conflict and affinity across human-likeness levels.

Predictor	Low Human-Likeness				Mid Human-Likeness				High Human-Likeness			
	<i>B</i>	<i>SE</i>	β	<i>t</i>	<i>B</i>	<i>SE</i>	β	<i>t</i>	<i>B</i>	<i>SE</i>	β	<i>t</i>
Constant	1.251	0.094	–	13.309 ***	2.441	0.215	–	11.353 ***	–0.822	0.314	–	–2.618 **
<i>S</i>	0.142	0.032	0.402	4.438 ***	0.421	0.061	0.584	6.902 ***	0.941	0.072	0.812	13.069 ***
<i>PC</i>	0.230	0.027	0.230	8.519 ***	–0.167	0.018	–0.167	–9.278 ***	–0.114	0.010	–0.114	–11.400 ***
<i>N</i>	453				244				303			
<i>R</i> ²	0.185				0.420				0.707			
<i>F</i>	51.240 ***				87.320 ***				362.150 ***			

Note: ** $p < 0.01$, *** $p < 0.001$. *B*—unstandardized coefficient; *SE*—standard error; β —standardized coefficient; *t*—*t* statistic; *N*—sample size; *R*²—coefficient of determination; *F*—overall model significance test.

To further examine how the association between the operational PC index and affinity varies across the human-likeness continuum, we divided the 1,000 samples into three groups according to the perceptual conflict index: low conflict ($PC < 0.083$), medium conflict ($0.083 \leq PC < 0.750$), and high conflict ($PC \geq 0.750$). We then fitted the relationship between human-likeness and affinity for each group, as shown in Figure 4.

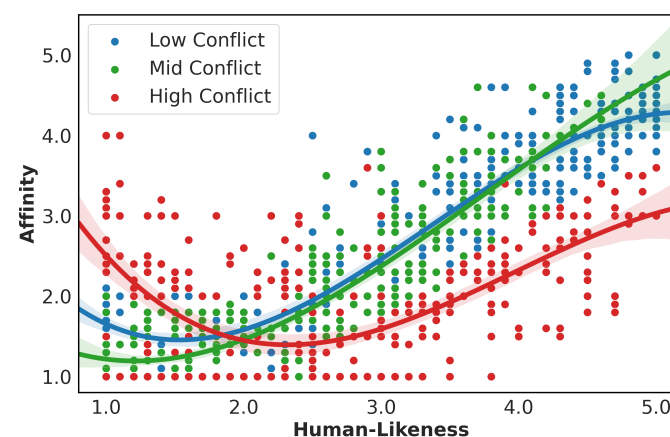


Figure 4. Association between operational PC levels and the relationship between human-likeness and affinity.

Specifically, in the low-conflict group, affinity stays at a moderate level in the low human-likeness range and then increases steadily as human-likeness rises, before gradually flattening at the highest levels. In the medium-conflict group, affinity starts from the lowest level among the three groups, but then increases sharply and eventually reaches the highest values in the high human-likeness range. In the high-conflict group, affinity is relatively high at the very low end of human-likeness, but then drops quickly and remains low across the low-to-mid human-likeness range, before showing only a limited recovery at higher human-likeness levels. Overall, these patterns show that the association between the operational PC index and affinity is not constant across the human-likeness continuum. At very low levels of human-likeness, high conflict is not immediately associated with lower affinity and may instead be interpreted as stylization. However, once characters move beyond the very low end of human-likeness, similar inconsistencies become less tolerable and are associated with a pronounced reduction in affinity across the low-to-mid human-likeness range. This pattern suggests that the valley-like decline in affinity is not driven by human-likeness alone, but is also associated with how cue-level inconsistency, as captured by the operational PC index, varies across levels of overall human-likeness.

Taken together, these analyses indicate that the operational perceptual conflict index is meaningfully associated with affinity ratings in the human-rating dataset used in this study. The hierarchical regression results show that this index accounts for additional variance in affinity ratings beyond overall human-likeness alone. The grouped regression results and visualization further indicate that this association varies across the human-likeness continuum and is related to the observed valley-like pattern in affinity ratings. Importantly, these findings should not be interpreted as directly measuring the full psychological process of perceptual conflict or as proving that UVE-PCoT internally reasons through this mechanism. Rather, they provide empirical motivation for adopting a perceptual-conflict-oriented structure in the subsequent UVE-PCoT perceptual decomposition. As presented in the following section, UVE-PCoT organizes the evaluation process by first establishing an overall human-likeness reference, then examining local-feature realism, evaluating whether these cues jointly support a coherent overall human-like impression or reveal cue-level inconsistency, and finally producing the affinity prediction.

4. Methods

To develop an interpretable framework for UVE evaluation, we draw on the perceptual-conflict-oriented perspective, which interprets UVE-related responses from the viewpoint of visual cue consistency. The overall methodology of this paper is illustrated in Figure 5, which comprises three main phases. First, in Phase (a), UVE Perceptual CoT Construction, we design the Perceptual Chain-of-Thought (UVE-PCoT) as a structured perceptual decomposition framework. This framework operationalizes the perceptual-conflict-oriented perspective into four staged evaluation components: overall human-likeness assessment, local-feature realism assessment, perceptual conflict detection, and affinity evaluation. Second, in Phase (b), Automated Perceptual CoT Data Annotation, we construct the UVE-R dataset using a multi-model collaborative correction pipeline that employs multiple VLMs for initial rationale generation, a VLM for fusion and consistency correction, and human experts for verification. This process produces image-grounded cue-level explanatory rationales consistent with the human rating scores in the dataset. These rationales follow the UVE-PCoT decomposition structure rather than directly reconstructing human observers' internal psychological reasoning. Finally, in Phase (c), UVE Perceptual CoT Model Training, we fine-tune three pre-trained vision-language models using LoRA-based supervised fine-tuning. The training pipeline itself follows a standard parameter-efficient fine-tuning paradigm. The contribution lies in using structured rationale supervision to

train each model to produce staged UVE evaluations and affinity predictions following the UVE-PCoT format.

4.1. UVE-PCoT: Constructing the Perceptual Chain-of-Thought for the Uncanny Valley Effect

Based on the perceptual-conflict-oriented perspective, which interprets the UVE from the perspective of cue-level consistency, we construct UVE-PCoT as a structured perceptual decomposition framework for interpretable UVE evaluation. This perspective suggests that affinity judgments are not determined by overall human-likeness alone, but may also be associated with whether local-feature realism supports or disrupts a coherent overall human-like impression. The empirical analyses in Section 3 further show that the human-rating data exhibit statistical patterns consistent with this perceptual-conflict-oriented perspective, providing empirical motivation for incorporating this cue-level interpretive structure into UVE-PCoT. Rather than treating affinity prediction as a direct mapping from visual appearance to a score, UVE-PCoT organizes UVE assessment into four sequential stages: overall human-likeness assessment, local-feature realism assessment, perceptual conflict detection, and affinity evaluation.

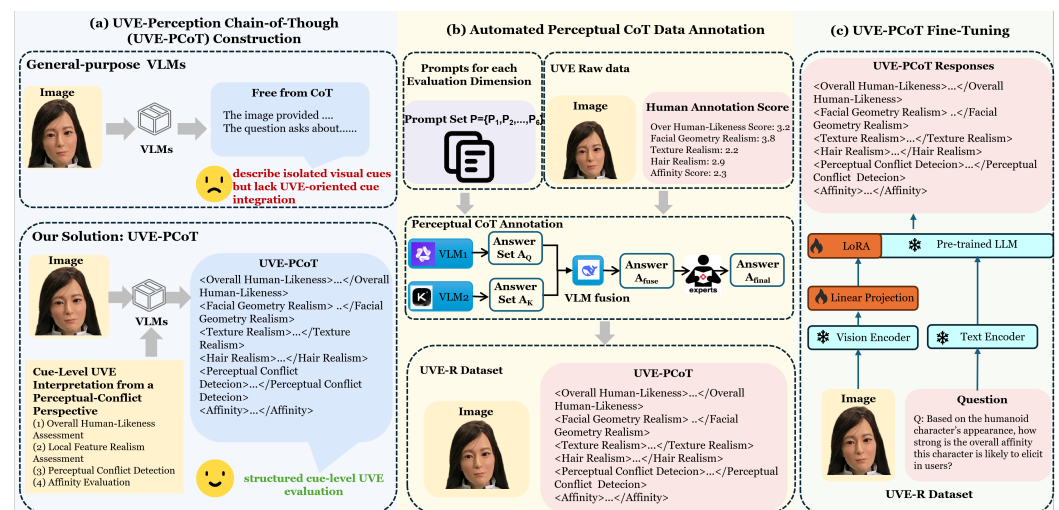


Figure 5. Overview of the proposed framework. (a) UVE-PCoT Construction: designing a staged perceptual decomposition structure guided by a perceptual-conflict-oriented perspective; (b) Perceptual CoT Data Annotation: constructing UVE-R through multi-model collaboration and expert verification to provide image-grounded, rating-consistent rationales; (c) UVE-PCoT Model Training: fine-tuning vision-language models with structured rationale supervision for interpretable UVE evaluation.

Importantly, UVE-PCoT does not aim to learn or reproduce the internal reasoning process of human observers. Instead, it defines a structured rationale format that makes the intermediate basis of UVE evaluation explicit and inspectable. In this structure, overall human-likeness provides the global reference, local-feature realism provides cue-level evidence, perceptual conflict examines whether local cues coherently support the overall impression, and affinity evaluation integrates these components into the final score. Accordingly, the UVE-R rationales used for supervision should be understood as image-grounded cue-level explanatory rationales consistent with the human rating scores in the dataset, rather than as direct evidence of human observers' internal psychological reasoning.

4.1.1. Stage 1: Overall Human-likeness Assessment

The first stage of UVE-PCoT assesses the overall human-likeness of the character and establishes a global reference against which subsequent local cues are interpreted. This score corresponds to the human-likeness dimension in the classical formulation of the UVE and provides the overall context for the following stages. Beyond serving as a numeric

rating, this stage describes the character's holistic human-like impression, including global properties that are difficult to reduce to isolated local features, such as vitality, naturalness, and lifelike presence. Accordingly, the model is trained to output both a qualitative description of the character's overall human-like impression and a quantitative human-likeness score on a 1–5 scale. This stage is supervised by UVE-R annotations constructed to be consistent with human ratings of overall human-likeness and serves as the reference for evaluating local-feature realism in the next stage.

4.1.2. Stage 2: Local-Feature Realism Assessment

Given the global reference established in Stage 1, UVE-PCoT next evaluates the realism of local facial cues. This stage focuses on three dimensions: facial geometry, texture, and hair. Facial geometry concerns structural attributes such as facial proportions, contours, and the arrangement of key components. Texture concerns surface appearance, including skin smoothness, material quality, wrinkles, pores, freckles, and other fine-grained details. Hair concerns hair-related cues, including head hair, eyebrows, eyelashes, facial hair, and their color, texture, and boundary integration. For each dimension, the model is trained to output both a qualitative description and a quantitative realism score on a 1–5 scale. These outputs are supervised by UVE-R annotations constructed to be consistent with the corresponding human ratings for the three local realism dimensions.

4.1.3. Stage 3: Perceptual Conflict Detection

In the third stage, UVE-PCoT analyzes perceptual conflict by relating the local-feature realism assessed in Stage 2 to the overall human-likeness reference established in Stage 1. The model is required to examine whether local cues remain mutually consistent and whether they jointly support a coherent overall human-like impression. This stage is designed to represent cases in which some local cues appear more or less realistic than expected under the overall impression, thereby suggesting weaker perceptual coherence. Unlike the rating-based stages, perceptual conflict detection is not supervised by a single independent numerical score. Instead, it is supervised through structured rationales that describe how inconsistencies among local cues, or between local cues and the overall impression, may be reflected in affinity evaluation.

The PC index constructed in Section 3 is not used as direct supervision for this stage. This is because the PC index is a rating-derived residual variance proxy computed from overall human-likeness and local-feature realism ratings, rather than an independently annotated perceptual conflict label. It is useful for empirically examining whether the human-rating data show statistical patterns consistent with the perceptual-conflict-oriented perspective in the current dataset, but it does not directly measure the full psychological process of perceptual conflict. Moreover, the rating dimensions from which the PC index is computed are already represented in the preceding stages, namely overall human-likeness assessment and local-feature realism assessment. Therefore, using the PC index as an additional scalar target would not provide an independent source of image-level perceptual conflict supervision, but could instead duplicate rating-derived information and encourage the model to fit a residual statistic rather than identify image-grounded cue-level inconsistencies.

Directly collecting perceptual conflict as a single user-provided score is also nontrivial. Unlike overall human-likeness, local-feature realism, and affinity, which can be collected as relatively direct perceptual or affective ratings, perceptual conflict is not a directly observable visual attribute, but a relational construct concerning whether different local features and their realism levels coherently support the overall human-like impression. If users were asked to assign a single perceptual conflict score directly, their responses could

easily be confounded with perceived abnormality, unfamiliarity, eeriness, or low affinity, and may depend strongly on how the task is explained. Accordingly, the numerical PC index and the language-based perceptual conflict stage are not intended to be formally equivalent. Section 3 provides empirical motivation for including a perceptual-conflict-oriented stage in the UVE-PCoT decomposition, whereas Stage 3 provides image-grounded, cue-level explanatory supervision. Specifically, Stage 3 is supervised through structured rationales that describe which visual cues appear inconsistent, how their realism levels differ, and how these inconsistencies are reflected in the coherence of the overall human-like impression. Future work could develop more precise ways to characterize perceptual conflict by combining fine-grained cue-level annotations, controlled manipulations of local-feature realism, response-time measures, eye-tracking, and other behavioral indicators. However, these behavioral or physiological indicators should not be interpreted as direct objective measurements of perceptual conflict as an independent psychological entity. Rather, they may provide complementary evidence about observable response patterns, such as attention allocation, processing difficulty, or response hesitation, when local visual cues are inconsistent. Such evidence would help examine whether the proposed cue-level conflict structure aligns with external manifestations of human rating behavior, while the present framework should still be understood as an interpretable computational tool for UVE evaluation rather than as a simulation of the full human perceptual process.

4.1.4. Stage 4: Affinity Evaluation

In the final stage, UVE-PCoT synthesizes the outputs of the previous stages to generate the final affinity rating, which corresponds to the dependent variable in UVE evaluation. This stage evaluates affinity by integrating overall human-likeness, local-feature realism, and cue-level inconsistencies identified from a perceptual-conflict-oriented perspective. Overall human-likeness provides the general basis for affinity, while inconsistencies among local cues may be associated with lower affinity when they fail to support a coherent overall human-like impression. As suggested by the empirical analysis, such inconsistencies may not be immediately associated with lower affinity in very low human-likeness contexts, where they can be interpreted as stylization. However, as overall human-likeness increases, similar inconsistencies become harder to reconcile with the character's overall appearance and are more likely to be associated with lower affinity. The model outputs a quantitative affinity score on a 1–5 scale together with a summary that links the preceding stages to the final judgment. This final stage is supervised by UVE-R annotations constructed to be consistent with human affinity ratings.

4.2. UVE-R Dataset: Structured Rationale Annotation Design for UVE Evaluation

To provide structured rationale supervision for UVE-PCoT, we build a structured rationale dataset, named UVE-R, on top of the humanoid character dataset of Li et al. [42]. While the original dataset provides images together with human ratings of overall human-likeness, facial geometry, texture, hair, and affinity, it does not include the staged rationale annotations needed to supervise UVE-PCoT outputs. To address this gap, we develop a multi-model collaborative correction pipeline for constructing structured rationale annotations. As shown in Figure 6, the pipeline consists of three stages: (a) initialization, (b) fusion, and (c) expert verification. Crucially, the construction process uses the original human rating scores as anchors to guide rationale construction and correction. Consequently, the resulting UVE-R rationales are image-grounded cue-level explanatory annotations that are consistent with the human rating scores in the dataset and reflect observable image features. They are not claimed to be direct evidence of the internal psychological reasoning processes of human observers. For clarity, the prompt illustrated in Figure 6 is an example template

for facial geometry realism assessment in humanoid characters. The full prompt library, including the templates for overall human-likeness, geometry realism, texture realism, hair realism, perceptual conflict detection, and affinity evaluation, is provided in Appendix C.

To align the annotation process with the staged structure of UVE-PCoT, we design a prompt library $P = \{P_1, P_2, \dots, P_6\}$, in which each prompt targets one component of the staged perceptual decomposition. These prompts are informed by the perceptual-conflict-oriented perspective and jointly cover the four stages of UVE-PCoT. Specifically, the six prompt types are defined as follows:

- Overall Human-Likeness (P_1): This prompt elicits a holistic judgment of the character's overall human-like appearance, considering the face as an integrated whole rather than as a collection of isolated features.
- Facial Geometry Realism (P_2): This prompt targets the realism of facial structure, including facial proportions, contours, and the arrangement of key components such as the eyes, nose, mouth, teeth, and eyebrows. It guides the model to describe and score how closely these structural attributes resemble those of a real human.
- Texture Realism (P_3): This prompt focuses on the realism of surface appearance, including skin color, smoothness, pores, wrinkles, freckles, moles, and other fine-grained texture cues. It guides the model to describe and score the realism of these texture-related cues.
- Hair Realism (P_4): This prompt evaluates the realism of hair-related cues, including head hair, eyebrows, eyelashes, facial hair, hairline transition, and their interaction with lighting and material appearance. It guides the model to describe and score how realistic these cues appear.
- Perceptual Conflict Detection (P_5): This prompt directs the model to compare local-cue realism with the overall human-likeness impression and identify inconsistencies among local cues that may suggest weaker perceptual coherence of the character.
- Affinity Evaluation (P_6): This prompt asks the model to generate an affinity evaluation by integrating the outputs of the previous steps, especially the relationship among overall human-likeness, local-feature realism, and cue-level inconsistencies.

Following this prompt design, the annotation pipeline proceeds in three stages.

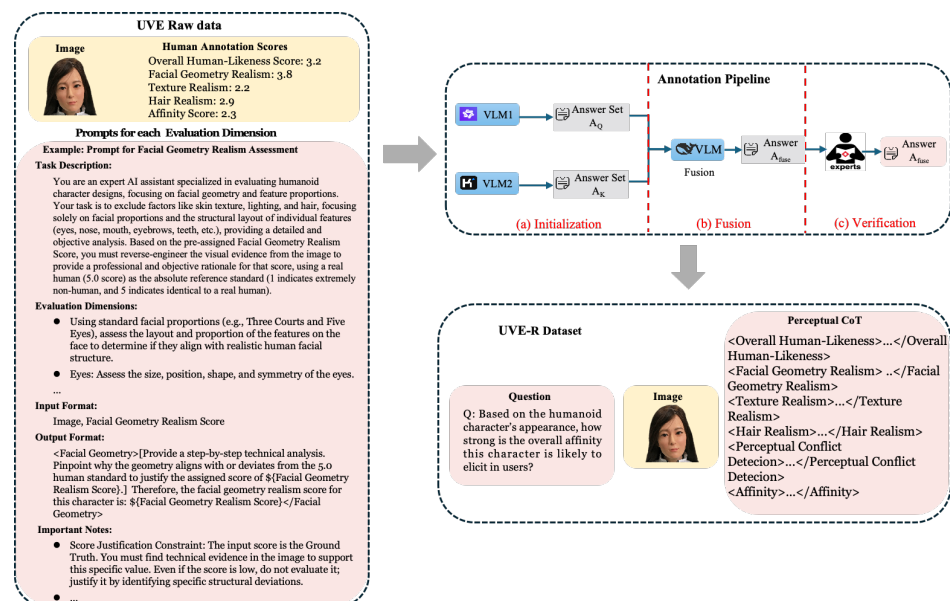


Figure 6. Overview of the UVE-R annotation pipeline for constructing structured rationale annotations as image-grounded cue-level explanatory rationales consistent with the human rating scores in the dataset.

4.2.1. Multimodal Collaborative Initialization

As discussed in the Introduction, VLMs with free-form CoT are not sufficiently reliable for direct UVE evaluation because they may identify visible anomalies but fail to consistently organize these cues within a UVE-oriented cue-integration structure for affinity evaluation. However, this limitation does not mean that VLMs fail to capture all perceptual dimensions relevant to UVE evaluation. Before using VLMs for annotation initialization, we conducted a preliminary dimension-level evaluation on the held-out test split used in Section 5, which contains 186 characters excluded from fine-tuning. This evaluation was used as an auxiliary feasibility check to examine whether candidate VLMs could provide reasonable initial descriptions and scores for the perceptual dimensions required by UVE-PCoT, rather than as a procedure for selecting fine-tuned checkpoints, tuning hyperparameters, or evaluating the final UVE-PCoT model. We evaluated the same candidate VLMs as those used in the experimental comparison, including GPT-5 [49], InternVL3.5, Qwen3.5-Plus, Gemini-3.1-Pro [50], and Kimi-K2.5 [51]. This evaluation focuses on the perceptual dimensions required for UVE-PCoT construction, namely overall human-likeness, geometry realism, texture realism, and hair realism. Affinity prediction is not included here because it serves as the final evaluation task and is reported separately in Section 5.

Table 3. Preliminary dimension-level evaluation of VLMs for candidate annotation initialization.

Perception Dimension	Model	Performance			
		MAE ↓	RMSE ↓	PC ↑	R ² ↑
Overall Human-Likeness	GPT-5	1.2120	1.4823	0.8202	0.2562
	InternVL3.5	0.4937	0.6312	0.9488	0.8583
	Qwen3.5-Plus	0.2778	0.3992	0.9752	0.9457
	Gemini-3.1-Pro	0.2467	0.3746	0.9768	0.9514
	Kimi-K2.5	0.4259	0.6054	0.9528	0.8752
Geometry Realism	GPT-5	2.0453	2.4792	0.0924	−1.0063
	InternVL3.5	0.4022	0.5382	0.9613	0.9028
	Qwen3.5-Plus	0.2373	0.3767	0.9822	0.9533
	Gemini-3.1-Pro	0.3114	0.4514	0.9688	0.9310
	Kimi-K2.5	0.4686	0.7081	0.9501	0.8351
Texture Realism	GPT-5	1.6402	2.1065	0.6097	−0.4562
	InternVL3.5	0.5061	0.7122	0.9494	0.8302
	Qwen3.5-Plus	0.2800	0.4964	0.9611	0.9187
	Gemini-3.1-Pro	0.3935	0.5985	0.9526	0.8824
	Kimi-K2.5	0.7762	1.1765	0.8498	0.5431
Hair Realism	GPT-5	0.3737	0.5857	0.9666	0.8890
	InternVL3.5	0.3022	0.4834	0.9714	0.9231
	Qwen3.5-Plus	0.2503	0.5326	0.9558	0.9075
	Gemini-3.1-Pro	0.5152	0.7225	0.9500	0.8305
	Kimi-K2.5	0.6568	0.9769	0.9148	0.6887

Note: PC denotes Pearson correlation with human ratings.

As shown in Table 3, the evaluated VLMs show different levels of agreement with human ratings across the four perceptual dimensions. For overall human-likeness, Gemini-3.1-Pro achieves the best performance, with the lowest MAE and RMSE and the highest correlation, followed closely by Qwen3.5-Plus. InternVL3.5 and Kimi-K2.5 also show high correlations, suggesting that most advanced VLMs can approximate holistic human-likeness judgments to some extent. For geometry realism, Qwen3.5-Plus performs best across all metrics, followed by Gemini-3.1-Pro and InternVL3.5, indicating that VLMs are relatively effective at recognizing structural facial cues such as proportions, contours, and component arrangement. For texture realism, Qwen3.5-Plus again achieves the strongest performance, while InternVL3.5 and Gemini-3.1-Pro remain competitive; however, the

larger errors of GPT-5 and Kimi-K2.5 suggest that skin-material realism is more difficult and model-dependent. For hair realism, GPT-5, InternVL3.5, and Qwen3.5-Plus all obtain high correlations, although their error metrics differ, indicating that hair-related cues can be detected by multiple VLMs but are not evaluated uniformly across models.

Overall, these results indicate that the limitations of VLMs in UVE affinity prediction do not stem solely from an inability to recognize basic perceptual dimensions. Advanced VLMs can capture several UVE-related perceptual components, including overall human-likeness and local-feature realism; however, their main limitation lies in integrating these cues for final affinity judgment. In free-form UVE evaluation, they may rely heavily on general definitions of the UVE or on isolated feature descriptions, without explicitly assessing whether the realism of different features jointly supports a coherent overall human-like impression. Nonetheless, these outputs can serve as useful initial candidates for dataset annotation, providing a practical starting point for subsequent fusion, correction, and verification in constructing UVE-R.

Given its consistently strong agreement with human ratings across the evaluated perceptual dimensions, Qwen3.5-Plus is used as the primary initialization model for generating candidate annotations. We further include Kimi-K2.5 as a secondary candidate source to avoid constructing all candidate annotations from a single VLM source. Although Kimi-K2.5 is not the strongest quantitative predictor in the preliminary evaluation, its outputs are not directly treated as final annotations; instead, they are used as additional candidates that are later fused, corrected, and verified by experts. Accordingly, we adopt a parallel initialization strategy using these two candidate sources. Specifically, two VLMs, Qwen3.5-Plus and Kimi-K2.5, denoted by f_Q and f_K , are used to process each input image together with the prompt template for a given UVE-PCoT component and, when available during annotation construction, its corresponding human rating.

For the i -th UVE-PCoT component, the multimodal generation process is written as

$$A_i = f_M(I, P_i, Y_i), \quad (4)$$

where $M \in \{K, Q\}$ denotes the VLM used in the initialization stage, I is the input image, P_i is the prompt template for the i -th component of UVE-PCoT, Y_i is the corresponding human rating used as a conditional anchor during annotation construction when available, and A_i is the generated textual answer for that component. Human ratings serve only as conditional anchors to guide the generation of rating-consistent rationales, and are not treated as targets to be re-predicted, nor included in the model input during fine-tuning or inference. In our setting, prompts P_1 – P_4 correspond to overall human-likeness, facial geometry realism, texture realism, and hair realism, respectively, and are accompanied by human ratings. The perceptual conflict prompt P_5 is not directly supervised by a single numerical score, while the affinity prompt P_6 uses human affinity ratings for UVE-R annotation.

Under the parallel initialization strategy, the same input configuration is processed by both models, yielding two candidate answers for each UVE-PCoT component:

$$A_{i,K} = f_K(I, P_i, Y_i), \quad A_{i,Q} = f_Q(I, P_i, Y_i), \quad (5)$$

which in turn produce two parallel answer sets:

$$A_K = \{A_{1,K}, A_{2,K}, \dots, A_{6,K}\}, \quad A_Q = \{A_{1,Q}, A_{2,Q}, \dots, A_{6,Q}\}. \quad (6)$$

4.2.2. Fusion and Semantic Aggregation

After generating the candidate answer sets A_K and A_Q , we introduce a fusion stage to integrate these parallel responses into a unified structured annotation A_{fuse} . In our

implementation, the fusion operator g is instantiated by DeepSeek-VL2 [52], which is used as a multimodal VLM to jointly process the input image, candidate rationales, available human ratings, and the fusion prompt. This multimodal design allows the fusion stage to verify whether candidate descriptions are visually grounded, rather than merely merging textual outputs. This stage is formalized through a fusion operator g :

$$A_{\text{fuse}} = g(A_K, A_Q, I, Y, P_{\text{fuse}}) \quad (7)$$

where I denotes the input image, Y denotes the available human ratings for the current sample, including overall human-likeness, facial geometry realism, texture realism, hair realism, and affinity, and P_{fuse} denotes a specialized fusion prompt for multi-source annotation reconciliation. The template of this fusion prompt is provided in Appendix C.

At this stage, only the image I and the available human ratings Y are treated as reference evidence for annotation construction, whereas A_K and A_Q are regarded as candidate structured rationales to be examined and corrected when necessary. The role of g is therefore not merely to merge two model outputs, but to compare the candidate rationales against the visual content and the available rating anchors, identify unsupported or hallucinated descriptions, resolve cross-model discrepancies, and produce a single structured annotation that remains consistent with both the image and the human ratings. In this way, the fusion stage consolidates and corrects the separate candidate rationales to obtain a more coherent and rating-consistent annotation for subsequent expert validation. Thus, the fusion stage uses human ratings as consistency anchors for constructing rating-consistent cue-level explanatory rationales, rather than as evidence for reconstructing the internal psychological reasoning processes of human observers.

4.2.3. Expert Validation

The final stage of the annotation pipeline uses expert validation to ensure the quality of the UVE-R dataset. We invited five experts with complementary experience in humanoid character evaluation to review the fused annotations A_{fuse} . Specifically, two experts had backgrounds in human-computer interaction or social robotics, with experience in evaluating anthropomorphic agents; two experts had backgrounds in computer graphics or digital human modeling, with experience in assessing visual realism and image-based annotation quality; and one expert had experience in perceptual evaluation of humanoid characters and UVE-related analysis. To reduce residual errors or hallucinations introduced during the fusion stage, the experts cross-check each fused annotation against the two raw model outputs A_K and A_Q , the original image I , and the available human ratings for the sample Y . The validation focuses on four aspects: whether the visual descriptions are grounded in the image, whether the numerical ratings are correctly reflected, whether the cue-level inconsistency analysis from the perceptual-conflict-oriented perspective is logically consistent with the local cues and overall impression, and whether unsupported or hallucinated descriptions are present. Formally, the final validated annotation is obtained as

$$A_{\text{final}} = v(A_K, A_Q, A_{\text{fuse}}, E \mid I, Y) \quad (8)$$

where E denotes expert knowledge, v denotes the refinement process, and Y denotes the available human ratings for the current sample.

Specifically, we adopt a reliability threshold of 80%: if at least four of the five experts independently agree that A_{fuse} correctly integrates the evidence from A_K and A_Q and remains consistent with both the visual content of I and the available human ratings, the annotation is retained. This threshold corresponds to a strong-majority agreement criterion rather than a simple majority, and is used as a practical quality-control rule for expert

validation. To reduce the potential influence of confirmation bias, the five experts first made their validation decisions independently before any group discussion or revision. For each initial fused draft, they judged whether the annotation was visually grounded, consistent with the human rating scores in the dataset, and logically coherent, rather than simply confirming the model-generated output. We further calculated inter-annotator agreement for these initial validation decisions. The Fleiss' Kappa across the five experts reached $\kappa = 0.74$, indicating substantial inter-expert agreement during the independent validation phase. If the threshold is not met, the case enters a collaborative revision stage, in which the experts discuss the identified problems and produce a consensus revision. The revised annotation is retained only after it satisfies the same grounding, rating-consistency, and coherence criteria.

To further examine the extent of expert intervention, we tracked the revision status of the fused drafts. The results show that approximately 38.5% of the initial fused drafts underwent substantial human revision, while 7.8% were completely rewritten from scratch after expert review. This indicates that the experts did not passively accept the model outputs, but actively corrected or replaced annotations when the fused rationales were insufficiently grounded, inconsistent with human ratings, or logically incomplete. During the collaborative revision stage, the experts jointly identified problematic parts, such as unsupported visual descriptions, rating-inconsistent statements, unclear cue-level inconsistency analysis, logically disconnected explanations, or hallucinated cues. The annotation was then revised by referring back to the original image, the candidate rationales, and the available human-rating anchors. Because the validation process was designed as an iterative correction procedure, annotations that failed the initial agreement criterion were revised and rechecked rather than directly used as final annotations. This verification step improves the logical consistency and visual faithfulness of the final annotation while further reducing residual model bias and hallucinations. As a result, the final UVE-R dataset provides structured rationale supervision for training UVE-PCoT in the form of cue-level explanatory rationales that are grounded in observable image features and consistent with the human rating scores in the dataset, rather than evidence of the internal psychological reasoning of human observers.

4.3. UVE-PCoT Model Fine-Tuning

To train models capable of producing structured UVE evaluation outputs, we fine-tune three pre-trained vision-language models, namely Qwen3-VL-4B, InternVL3.5-4B, and LLaVA-1.5-7B, using the proposed UVE-R dataset. The training pipeline follows a standard LoRA-based supervised fine-tuning paradigm. The methodological contribution does not lie in introducing a new optimization algorithm or a new model architecture, but in using structured rationale annotations as supervision. These rationales are constructed with the human rating scores in the dataset as anchors during UVE-R construction and are therefore intended as rating-consistent, image-grounded cue-level explanatory rationales rather than direct evidence of human observers' internal psychological reasoning. Unlike conventional supervision that maps an input image directly to a final affinity score, UVE-R provides structured annotations covering overall human-likeness, local-feature realism, perceptual conflict detection, and final affinity evaluation. This supervision enables each fine-tuned model to output structured UVE assessments together with affinity predictions.

To avoid label leakage, the input during supervised fine-tuning contains only the image I and the task query Q . Human ratings are not provided in the input prompt. Instead, the staged rationales and their corresponding rating values are included only in the target output sequence Y . Thus, the model is trained to generate the UVE-PCoT evaluation from visual input and task instruction, rather than being conditioned on ground-

truth scores. Specifically, for each sample, the final validated annotation is converted into a standardized text sequence $Y = y_1, y_2, \dots, y_L$. Given the input image I and a task query Q , the model is trained with the standard cross-entropy (CE) loss:

$$\mathcal{L}_{\text{CE}} = - \sum_{t=1}^L \log p(y_t | y_{<t}, I, Q), \quad (9)$$

where $p(y_t | y_{<t}, I, Q)$ denotes the conditional probability of predicting the target token y_t given the preceding tokens, the image, and the task query.

To ensure a rigorous comparative analysis, a core set of training settings is kept consistent across all evaluated models (Qwen3-VL-4B, InternVL3.5-4B, and LLaVA-1.5-7B). Specifically, we employ Low-Rank Adaptation (LoRA) [53] with rank $r = 32$, $\alpha = 64$, and a dropout rate of 0.05. Crucially, across all backbones, the vision encoders remain strictly frozen, while LoRA adapters are consistently applied to the linear projection layer connecting the vision encoder to the language model. All training is conducted on four NVIDIA RTX 3090 GPUs with gradient checkpointing enabled. We universally utilize the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, weight decay 0.1) and maintain a consistent effective batch size of 16 (per-device batch size of 1 with 4 gradient accumulation steps). The random seed is fixed to 42 for all training and inference procedures.

Conversely, to accommodate architectural variations and preserve the distinct convergence dynamics of each backbone, several settings are model-specific. For Qwen3-VL-4B and InternVL3.5-4B, LoRA adapters are applied to the language model's `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules, and training spans 10 epochs using a cosine learning-rate schedule decaying from 1×10^{-5} to 5×10^{-9} (warm-up ratio 0.03). Regarding preprocessing, Qwen3-VL-4B uses dynamic resolution (`min_pixels` = 3136, `max_pixels` = 200,704) with a maximum sequence length of 8192, whereas InternVL3.5-4B images are resized to 448×448 with a 4096 sequence length. For LLaVA-1.5-7B, images are resized to 336×336 with a sequence length of 4096. Following established community best practices for LLaVA, its LoRA adapters are applied only to the `q_proj` and `v_proj` layers of the language model (alongside its linear projection layer). Furthermore, to prevent overfitting given its specific architecture, the learning rate for LLaVA-1.5-7B is adjusted to 2×10^{-4} and training is completed in 3 epochs.

The above hyperparameters were fixed before final evaluation. No separate validation set was used for systematic hyperparameter search. Instead, the training configuration was determined based on official and community-recommended fine-tuning practices for each backbone, common LoRA fine-tuning experience, and training stability observations on the training set. During training, we monitored the training loss and generation stability only to ensure normal convergence and valid output formatting. The held-out test set was strictly reserved for final evaluation and was not used for hyperparameter tuning, model selection, checkpoint selection, or early stopping.

During inference, models receive the exact input format used in training (image and task query only), with no human annotation scores, intermediate ratings, or cue labels provided. Deterministic greedy decoding is applied (`temperature` = 0.0, `do_sample` = False, `num_beams` = 1, `max_new_tokens` = 2048) under `torch.inference_mode()`. For all ablation variants based on the same backbone, the experimental condition is the sole independent variable. All other training and evaluation settings, including train/test splits, random seeds, and parsing rules, are kept strictly constant. The image-only fine-tuning baselines follow the same backbone models, LoRA settings, optimizer, training epochs, decoding strategy, train/test split, random seed, and score parsing rules as the corresponding full UVE-PCoT models. They differ only in the supervision target, as they are trained to directly output the final affinity score without staged rationale supervision.

5. Experiments

5.1. Experimental Design Overview

This section evaluates UVE-PCoT through a multi-level experimental design. Since UVE-PCoT is not intended to serve merely as an image-to-score prediction model, but as an interpretable assessment framework based on structured perceptual decomposition, its evaluation should examine not only whether the final affinity scores are accurately predicted, but also whether the proposed structure contributes to model performance and whether the generated explanations provide meaningful visual evidence. Therefore, the experiments are organized around three complementary levels: outcome-level effectiveness, component-level contribution, and explanation-level validity.

First, we conduct an affinity prediction experiment to provide outcome-level validation of UVE-PCoT. This experiment examines whether UVE-PCoT improves affinity prediction performance compared with alternative evaluation settings. Specifically, we compare four conditions: (1) general-purpose VLMs with free-form UVE explanations and no structured prompting, (2) the same VLMs prompted with the UVE-PCoT template without fine-tuning, (3) fine-tuned open-source VLMs with image-only direct prediction of affinity scores, and (4) fine-tuned open-source VLMs with full UVE-PCoT structured rationale supervision. This design separates the role of structured UVE-PCoT prompting from free-form explanation, and the role of structured rationale supervision from standard image-only fine-tuning. It also examines whether the benefit of structured supervision is observable under the same experimental condition across multiple open-source backbone models, including Qwen3-VL-4B, InternVL3.5-4B, and LLaVA-1.5-7B.

Second, we conduct ablation and control experiments to provide component-level evidence for the contribution of the UVE-PCoT structured perceptual decomposition. Starting from the complete four-stage structure, we remove the perceptual conflict stage, the local-feature realism stage, and the overall human-likeness assessment stage in separate settings. We also test four control settings. In the hidden-CoT setting, structured rationale supervision is retained during training, but intermediate rationales are not explicitly output during inference. In the shuffled-order setting, the order of the reasoning stages is changed. In the generic-rationale setting, image-specific rationales are replaced with generic explanations. In the conflict-corrupted setting, perceptual-conflict rationales are replaced with inconsistent explanations. These controls examine whether the observed performance gain is associated with valid, image-grounded structured rationales and the intended stage order, rather than with superficial textual templates, longer output sequences, or the mere presence of textual rationales.

Third, we evaluate the model's ability to identify visual cues associated with users' affinity judgments, in order to provide explanation-level validation of UVE-PCoT. Under the proposed cue-coding protocol, this experiment examines whether the cues identified in model-generated rationales are consistent with human-perceived cue references associated with affinity judgments, rather than merely producing plausible textual descriptions. This provides additional evidence that UVE-PCoT can generate cue-level explanations supported by visual evidence, without assuming that the model reproduces the perceptual processes of human observers.

Together, these three experiments form a complementary validation framework for the method proposed in Section 4. The affinity prediction experiment examines whether UVE-PCoT improves the final affinity assessment; the ablation and control experiments examine whether the proposed staged structure and rationale supervision contribute to performance; and the cue identification experiment examines whether the generated explanations provide visually grounded cue-level explanatory evidence. In this way, the experimental design jointly evaluates the predictive accuracy, structural contribution, and

interpretability of UVE-PCoT. Table 4 summarizes the core question, validation focus, and evidence provided by each experiment.

Table 4. Overview of the experimental design.

Validation Level	Experiment	Focus and Evidence
Outcome-level effectiveness	Quantitative evaluation of affinity prediction	Core question: Whether UVE-PCoT improves final affinity assessment across different model groups and fine-tuning settings. Main comparison: General-purpose VLMs with free-form explanations, general-purpose VLMs with the UVE-PCoT prompting template but without fine-tuning, image-only fine-tuned baselines, and full UVE-PCoT fine-tuned models. The fine-tuning experiments are conducted on three open-source backbones: Qwen3-VL-4B, InternVL3.5-4B, and LLaVA-1.5-7B. Metrics: MAE, RMSE, Pearson correlation, and R^2.
Component-level contribution	Ablation and control study on the staged structure	Core question: Whether the staged perceptual decomposition, stage order, and image-grounded rationale content contribute to affinity prediction performance. Main comparison: Stage-removal variants and control settings, including Hidden-CoT, Shuffled-Order, Generic-Rationale, and Conflict-Corrupted variants. Metrics: MAE, RMSE, Pearson correlation, R^2, and significance tests.
Explanation-level validity	Cue identification and interpretability	Core question: Whether the visual cues identified in generated rationales are consistent with human-perceived cues associated with affinity judgments. Main comparison: Alignment between cues identified in model-generated rationales and human-perceived cue references coded under the proposed cue-coding protocol, compared across UVE-PCoT and general-purpose VLM baselines. Metrics: Precision, recall, and F1-score.

5.2. Experiment 1: Quantitative Evaluation of Affinity Prediction

We first evaluate UVE-PCoT on the primary task of affinity prediction using human-annotated affinity ratings. The evaluation includes both general-purpose VLM baselines and fine-tuned open-source VLM backbones. The general-purpose VLMs include GPT-5, InternVL3.5, Qwen3.5-Plus, Gemini-3.1-Pro, and Kimi-K2.5. For fine-tuning-based comparisons, we use three representative open-source VLM backbones, namely Qwen3-VL-4B, InternVL3.5-4B, and LLaVA-1.5-7B. To ensure a fair comparison, we adopt stratified sampling based on character categories to divide the UVE-R dataset into a training set and a test set. The training set contains 814 characters, and the test set contains 186 characters that are excluded from fine-tuning.

We compare four groups of settings in this experiment. First, we evaluate general VLMs under free-form UVE explanations, where the models are asked to predict affinity without being constrained by the staged evaluation structure of UVE-PCoT. Second, we evaluate the same VLMs with the UVE-PCoT prompting template but without fine-tuning, in order to assess the effect of structured perceptual prompting alone. Third, we evaluate image-only fine-tuned baselines, in which the three open-source backbones are fine-tuned to directly predict affinity from the input image without generating staged rationales. This setting isolates the effect of structured rationale supervision by keeping the data split, model backbone, image preprocessing, LoRA configuration, optimization settings, random seed, and evaluation protocol consistent with the full UVE-PCoT setting. For each respective backbone, the image-only baselines use the same training hyperparameters as those reported in Tables A1 and A2, except that the target output is changed from staged rationales plus the final affinity score to the final affinity score only. During inference, the affinity score is extracted from the final model output for all fine-tuned settings. We note that because the image-only target sequence is substantially shorter than the structured rationale sequence, the same number of training epochs may represent a different effective training regime for the two settings. However, holding hyperparameters constant remains the standard approach for isolating the effect of a single experimental factor, consistent with the ablation methodology adopted in Section 5.3. Fourth, we evaluate the same three backbones fine-tuned with the full UVE-PCoT-based structured rationale supervision.

This comparison allows us to distinguish the effect of ordinary image-to-score fine-tuning from the effect of UVE-PCoT-based structured rationale supervision, while also examining whether the benefit is consistent across different model architectures. All evaluations are conducted on the test set. We measure prediction performance using four standard metrics [54,55]:

- Pearson correlation coefficient (PC), which measures the linear correlation between predicted and human-rated affinity scores.
- Coefficient of determination (R^2), which indicates the proportion of variance in human ratings explained by the model.
- Mean absolute error (MAE), which measures the average magnitude of prediction errors.
- Root mean square error (RMSE), which places greater weight on larger prediction errors.

Table 5 reports the quantitative results of affinity prediction under different prompting and fine-tuning settings. Under the free-form setting, model performance is highly uneven. InternVL3.5 achieves the strongest result in this group, with $R^2 = 0.2566$ and $PC = 0.5911$, whereas the remaining models perform substantially worse, including negative R^2 values for GPT-5, Qwen3.5-Plus, Gemini-3.1-Pro, and Kimi-K2.5. These results suggest that unconstrained free-form multimodal explanations are generally insufficient for reliable affinity prediction in the context of UVE evaluation.

Table 5. Quantitative comparison of affinity prediction performance across free-form explanations, UVE-PCoT prompting, image-only fine-tuning (FT), and UVE-PCoT structured rationale supervision.

Model	Setting	Affinity			
		$R^2 \uparrow$	PC $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
GPT-5	Free	−0.7305	−0.0659	1.1816	1.3836
InternVL3.5	Free	0.2566	0.5911	0.6832	0.9069
Qwen3.5-Plus	Free	−0.4887	0.1768	1.0765	1.2875
Gemini-3.1-Pro	Free	−0.8897	0.3461	1.2201	1.4437
Kimi-K2.5	Free	−0.1084	0.4089	0.9184	1.1073
GPT-5	+PCoT Prompt	0.1073	0.6457	0.7470	0.9068
InternVL3.5	+PCoT Prompt	0.5842	0.7857	0.5163	0.6800
Qwen3.5-Plus	+PCoT Prompt	0.5598	0.7568	0.5484	0.6941
Gemini-3.1-Pro	+PCoT Prompt	0.5927	0.7939	0.5085	0.6714
Kimi-K2.5	+PCoT Prompt	0.1488	0.5178	0.8151	0.9704
Qwen3-VL-4B	Image-only FT	0.5450	0.7422	0.4904	0.6430
InternVL3.5-4B	Image-only FT	0.5796	0.7648	0.4687	0.6215
LLaVA-1.5-7B	Image-only FT	0.3729	0.6284	0.6127	0.7685
Qwen3-VL-4B	UVE-PCoT FT	0.8088	0.9046	0.2124	0.4609
InternVL3.5-4B	UVE-PCoT FT	0.7854	0.8912	0.2316	0.4827
LLaVA-1.5-7B	UVE-PCoT FT	0.6418	0.7935	0.3549	0.6078

When the same models are prompted with the UVE-PCoT template, performance improves consistently across all tested VLMs. For example, GPT-5 improves from $R^2 = -0.7305$ and $PC = -0.0659$ in the free-form setting to $R^2 = 0.1073$ and $PC = 0.6457$ under UVE-PCoT prompting. InternVL3.5 further improves from $R^2 = 0.2566$ to 0.5842, while MAE decreases from 0.6832 to 0.5163. Similar gains are observed for Qwen3.5-Plus, Gemini-3.1-Pro, and Kimi-K2.5. Among the prompting-only baselines, Gemini-3.1-Pro achieves the best overall result, with $R^2 = 0.5927$, $PC = 0.7939$, $MAE = 0.5085$, and $RMSE = 0.6714$. These results indicate that structured perceptual prompting substantially improves affinity prediction relative to free-form UVE explanations.

We further compare image-only direct prediction fine tuning with full UVE-PCoT-based structured rationale supervision across three open-source backbones. In the image-

only setting, the models are fine tuned to predict affinity directly from the input image without generating staged rationales. This setting improves over most free-form VLM baselines, showing that task-specific fine tuning itself is beneficial. Among the image-only fine-tuned models, InternVL3.5-4B achieves the best result, with $R^2 = 0.5796$, PC = 0.7648, MAE = 0.4687, and RMSE = 0.6215, followed by Qwen3-VL-4B with $R^2 = 0.5450$ and PC = 0.7422. However, these results are still clearly lower than those obtained with full UVE-PCoT supervision.

Under full UVE-PCoT structured rationale supervision, all three fine-tuned backbones achieve stronger performance than their image-only counterparts. Qwen3-VL-4B+UVE-PCoT obtains the best overall result, with $R^2 = 0.8088$, PC = 0.9046, MAE = 0.2124, and RMSE = 0.4609. InternVL3.5-4B+UVE-PCoT also achieves competitive performance, with $R^2 = 0.7854$, PC = 0.8912, MAE = 0.2316, and RMSE = 0.4827. LLaVA-1.5-7B+UVE-PCoT improves over its image-only counterpart as well, increasing R^2 from 0.3729 to 0.6418 and PC from 0.6284 to 0.7935. These results indicate that the performance gain is not attributable to fine tuning alone. Instead, UVE-PCoT supervision provides additional benefits by guiding the model to organize affinity prediction through a structured perceptual decomposition involving overall human-likeness, local-feature realism, perceptual conflict detection, and final affinity evaluation.

Taken together, the results provide complementary evidence across prompting and fine-tuning settings. Compared with free-form explanations, UVE-PCoT prompting improves affinity prediction by explicitly structuring the evaluation process. Compared with image-only fine-tuning, which learns a direct mapping from the input image to the affinity score, UVE-PCoT structured rationale supervision further improves performance by providing intermediate perceptual annotations. This pattern suggests that affinity prediction in UVE evaluation benefits not only from general vision-language knowledge or task-specific fine tuning, but also from organizing perceptual observations around the relationship among overall human-likeness, local-feature realism, and cue-level inconsistency from a perceptual-conflict-oriented perspective. Therefore, the advantage of UVE-PCoT lies in advancing affinity prediction from direct image-to-score mapping toward structured and interpretable UVE evaluation.

To better understand why several VLMs produce negative R^2 values under the free-form setting, we conduct an error analysis on the full test set ($N = 186$). In our experiment, if a model output does not contain a valid 1–5 affinity score or is otherwise unparsable, the model is re-queried with a clarification instruction. Therefore, this analysis focuses on the final valid numerical predictions and excludes parsing failures or out-of-range responses. We compute four indicators to examine whether the negative R^2 values mainly reflect score calibration issues. For each image i , let $\hat{y}_i$ denote the model-predicted affinity score (1–5) and y_i the corresponding human rating. The four indicators are defined as follows:

- Bias: the average prediction error $\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)$. Positive values indicate systematic overestimation, while negative values indicate underestimation. This measures whether the model tends to produce scores that are generally too high or too low.
- SD ratio: the standard deviation of model predictions divided by the standard deviation of human ratings. A ratio below 1 indicates that the model's predictions are compressed and less sensitive to differences among images, whereas a ratio above 1 indicates overly dispersed predictions.
- Large-error rate: the proportion of samples for which $|\hat{y}_i - y_i| \geq 1$. This indicates how often the model produces substantial errors on the 1–5 scale.
- Low-affinity overestimation rate: focusing on the 20% of images with the lowest human ratings, the proportion for which the model predicts a score above the

neutral point of 3. This directly measures whether the model fails to recognize low-affinity characters.

As shown in Table 6, the free-form predictions show clear evidence of poor score calibration. All four models exhibit positive bias, with an average bias of 0.86, indicating that they generally assign higher affinity scores than human raters. This overestimation is particularly pronounced for Gemini-3.1-Pro and GPT-5, whose biases reach 0.94 and 0.89, respectively. In addition, the SD ratios are consistently below 1.0, with an average value of 0.57, suggesting that the predicted scores are compressed into a narrower range than human ratings. This indicates that free-form VLMs tend to produce moderate affinity scores and are less sensitive to image-level variation in human affinity judgments.

Table 6. Error-pattern analysis of valid free-form affinity predictions for models with negative R^2 on the test set.

Model	Bias	SD Ratio	Large Err. (%)	Low Over. (%)
GPT-5	0.89	0.52	72.6	18.4
Qwen3.5-Plus	0.76	0.61	64.5	13.2
Gemini-3.1-Pro	0.94	0.55	75.3	21.1
Kimi-K2.5	0.84	0.60	53.7	8.0
Average	0.86	0.57	66.5	15.2

The large-error results further clarify the source of the negative R^2 values. On average, 66.5% of the predictions deviate from human ratings by at least one scale point, showing that substantial absolute errors occur frequently. The low-affinity overestimation rate is lower than the large-error rate but remains non-negligible, with an average of 15.2%. This indicates that a subset of low-affinity samples is still assigned above-neutral scores, although the dominant error pattern is broader score miscalibration rather than low-affinity overestimation alone. This pattern also explains why some models may show non-zero Pearson correlations while obtaining negative R^2 : they may preserve limited ranking information, but their absolute scores are poorly calibrated to the human 1–5 affinity scale.

Overall, the poor performance of free-form VLMs, as indicated by negative R^2 values, suggests systematic overestimation, compressed prediction ranges, and insufficient sensitivity to low-affinity UVE samples. Specifically, these models tend to rely on surface-level definitions of the UVE, attributing low affinity predominantly to characters with high human-likeness and assuming that the valley occurs primarily in the high-likeness range. However, empirical evidence from the dataset used in this study indicates that the UVE emerges at low-to-mid human-likeness levels (approximately 1.7 on a 1–5 scale), rather than in the high-likeness region [42]. This discrepancy does not necessarily indicate that general VLMs lack the ability to recognize visible perceptual cues. Rather, it suggests that they have difficulty integrating separately identified cues into a coherent affinity evaluation. The perceptual-conflict-oriented perspective provides a cue-level interpretation of this issue by emphasizing whether local-feature realism jointly supports a coherent overall human-like impression and how such cue-level inconsistency is reflected in affinity judgments. This is precisely the motivation for UVE-PCoT, which organizes affinity prediction through overall human-likeness assessment, local-feature realism assessment, perceptual conflict detection, and final affinity evaluation.

5.3. Experiment 2: Ablation Study on the Staged Structure of UVE-PCoT

To examine the contribution of the staged evaluation structure and structured rationale supervision in UVE-PCoT, we conduct a set of ablation and control experiments using the full UVE-PCoT model as the reference configuration. All variants in this experiment are

implemented using Qwen3-VL-4B as the fixed backbone, so that the effect of modifying the UVE-PCoT structure can be isolated without introducing additional variation from different model architectures. Cross-backbone generality is examined separately in Experiment 1. Therefore, Experiment 2 should be interpreted as an internal ablation analysis on the held-out test set of the current dataset, rather than as cross-dataset validation under different rating distributions or covariance structures. The full UVE-PCoT model performs affinity prediction through the complete staged evaluation structure, including overall human-likeness assessment, local-feature realism assessment, perceptual conflict detection, and final affinity evaluation.

We first construct three stage-removal variants to evaluate the contribution of individual components in the UVE-PCoT structure. The first variant, w/o Perceptual Conflict, removes the perceptual conflict detection stage while retaining overall human-likeness assessment and local-feature realism assessment. This variant tests whether explicitly examining the consistency between local cues and the overall impression contributes to affinity prediction. The second variant, w/o Local-Feature Realism, removes the local-feature realism stage, so that the model no longer explicitly evaluates geometry, texture, and hair realism before predicting affinity. This variant tests whether cue-level realism assessment contributes to reliable UVE evaluation. The third variant, w/o Overall Human-Likeness, removes the overall human-likeness assessment stage, thereby eliminating the explicit holistic reference used to interpret local realism and perceptual conflict. This variant tests whether the global human-like impression provides an important reference for the subsequent staged evaluation.

We further include four control settings to examine whether the performance gain depends on valid, image-grounded structured rationales rather than on superficial textual templates or shortcut patterns. The Shuffled-order UVE-PCoT setting retains the same stage content but changes the intended stage order, testing whether the sequence of the staged evaluation is meaningful. The Conflict-Corrupted UVE-PCoT setting corrupts the perceptual conflict descriptions, testing whether incorrect conflict rationales weaken prediction performance. The Generic Rationale setting replaces image-specific rationales with generic descriptions, testing whether the model benefits from image-grounded rationale content rather than from template-like text. Finally, the hidden-CoT setting keeps structured rationale supervision during training but suppresses explicit intermediate rationale output during inference, so that only the final affinity prediction is generated. This setting helps distinguish the predictive benefit of structured supervision from the interpretability benefit of explicitly generated rationales.

For fair comparison, the ablation study uses the same data split and evaluation metrics as Experiment 1. All configurations are trained and evaluated on the same training and test sets, and prediction performance is measured by PC, R^2 , MAE, and RMSE. Since the purpose of this experiment is to examine the role of the staged perceptual structure itself, all training conditions are kept identical across configurations except for the staged components removed or altered in each variant.

Table 7 reports the ablation and control results. Overall, the full UVE-PCoT model achieves the best performance on all four metrics, with $R^2 = 0.8088$, PC = 0.9046, MAE = 0.2124, and RMSE = 0.4609. Performance decreases when key stages are removed, when the intended stage order is disrupted, or when the rationale content is corrupted or made generic. These results indicate that the UVE-PCoT structure is not merely a formal convention, but provides functionally meaningful structured rationale supervision for affinity prediction.

Removing the perceptual conflict stage leads to a clear decline in performance, with R^2 dropping from 0.8088 to 0.7580, PC from 0.9046 to 0.8670, MAE increasing from 0.2124

to 0.3360, and RMSE from 0.4609 to 0.5480. Although this variant still retains overall human-likeness assessment and local-feature realism assessment, it no longer explicitly examines whether local cues remain consistent with the overall impression. This result suggests that perceptual conflict detection is an important intermediate component that helps connect overall impression and local-cue analysis to final affinity evaluation.

Table 7. Ablation and control experiments of UVE-PCoT with Qwen3-VL-4B as the fixed backbone.

Model Variant	$R^2 \uparrow$	PC $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
Full UVE-PCoT	0.8088	0.9046	0.2124	0.4609
w/o Perceptual Conflict	0.7580	0.8670	0.3360	0.5480
w/o Local-Feature Realism	0.6550	0.8080	0.4210	0.5860
w/o Overall Human-Likeness	0.6184	0.7815	0.4526	0.6118
Shuffled-order UVE-PCoT	0.7814	0.8843	0.2736	0.5057
Conflict-Corrupted UVE-PCoT	0.5569	0.7476	0.4759	0.6374
Generic Rationale	0.5768	0.7596	0.4673	0.6271
Hidden-CoT	0.7962	0.8979	0.2297	0.4748

A larger performance drop is observed when the local-feature realism stage is further removed. In this case, R^2 decreases to 0.6550, PC to 0.8080, MAE rises to 0.4210, and RMSE to 0.5860. This result indicates that local-feature realism assessment provides essential fine-grained visual evidence for affinity prediction. Without explicit analysis of facial geometry, texture, and hair, the model retains only a coarser basis for evaluation and loses access to the local perceptual information required for cue-level inconsistency analysis and affinity estimation. Performance deteriorates further after removing the overall human-likeness stage. Under this configuration, R^2 falls to 0.6184, PC to 0.7815, MAE rises to 0.4526, and RMSE to 0.6118. This result suggests that overall human-likeness is not a secondary component, but a global reference for the staged evaluation structure. Without an initial judgment of how human-like the character appears as a whole, the model loses the perceptual context needed to interpret whether local cues support or undermine the character's overall impression.

The control settings further clarify why the full UVE-PCoT structure is effective. The shuffled-order variant retains the same stage content but changes the intended order of the staged evaluation. Its performance remains relatively strong, with $R^2 = 0.7814$, PC = 0.8843, MAE = 0.2736, and RMSE = 0.5057, but is still lower than the full model. This indicates that the performance gain is not only due to including more intermediate content; organizing the stages in the intended perceptual order also contributes to stable prediction.

The conflict-corrupted and generic-rationale settings lead to much larger performance drops. When perceptual conflict descriptions are corrupted, R^2 decreases to 0.5569, PC to 0.7476, MAE increases to 0.4759, and RMSE to 0.6374. Similarly, replacing image-specific rationales with generic rationales results in $R^2 = 0.5768$, PC = 0.7596, MAE = 0.4673, and RMSE = 0.6271. These results show that the model does not benefit simply from the presence of additional explanatory text. Instead, the rationales must be valid and image grounded. Incorrect conflict descriptions or generic template-like rationales weaken prediction substantially, supporting the claim that UVE-PCoT relies on meaningful structured rationale supervision rather than superficial textual patterns.

The Hidden-CoT setting achieves performance close to the full model, with $R^2 = 0.7962$, PC = 0.8979, MAE = 0.2297, and RMSE = 0.4748. This result suggests that structured rationale supervision remains useful even when the intermediate rationale stages are not explicitly shown during inference. However, the Hidden-CoT setting still performs slightly worse than the full UVE-PCoT model and does not provide explicit intermediate expla-

nations. Therefore, while hidden structured supervision can support prediction, the full UVE-PCoT output better preserves both predictive accuracy and interpretability.

To assess the statistical reliability of the ablation results, we conducted paired two-tailed *t*-tests on the per-sample absolute prediction errors between the full UVE-PCoT model and each variant over the same test set ($N = 186$). To control the family-wise error rate, we applied Holm correction for multiple comparisons. The full UVE-PCoT model significantly outperformed all variants except Hidden-CoT. Specifically, removing overall human-likeness ($t(185) = 12.34$, adjusted $p < 0.001$), removing local-feature realism ($t(185) = 9.87$, adjusted $p < 0.001$), and removing perceptual conflict ($t(185) = 8.42$, adjusted $p < 0.001$) each led to significant increases in prediction error. These results suggest that the three perceptual stages provide complementary supervision for affinity prediction. Overall human-likeness offers a global perceptual basis, local-feature realism provides fine-grained visual evidence, and perceptual conflict detection provides a structured step for examining whether local cues jointly support the overall impression before final affinity evaluation. The shuffled-order variant also showed significant degradation ($t(185) = 3.56$, adjusted $p < 0.001$), indicating that the sequential organization of evaluation stages contributes to model performance rather than serving only as a formatting convention. Both the conflict-corrupted rationale ($t(185) = 14.21$, adjusted $p < 0.001$) and the generic-rationale variant ($t(185) = 13.05$, adjusted $p < 0.001$) resulted in substantial performance drops, indicating that the benefit of UVE-PCoT cannot be explained by longer text length alone, but depends on task-relevant and semantically consistent rationale content.

The Hidden-CoT variant, which receives the same staged supervision during training but outputs only the final affinity prediction without explicit rationales, achieved numerically close performance to the full model (MAE 0.2297 vs. 0.2124, RMSE 0.4748 vs. 0.4609, PC 0.8979 vs. 0.9046, and R^2 0.7962 vs. 0.8088). The difference was not statistically significant ($t(185) = 1.23$, $p = 0.22$). This result suggests that structured staged supervision can preserve much of the predictive benefit even when intermediate rationales are not explicitly exposed in the final output. However, the purpose of UVE-PCoT extends beyond maximizing predictive accuracy. Unlike Hidden-CoT, the full UVE-PCoT setting provides explicit, human-readable intermediate explanations, enabling users to inspect how overall human-likeness, local-feature realism, and cue-level inconsistency analysis are organized before the final affinity judgment. This transparency is particularly valuable for design-oriented UVE evaluation, where interpretable cue-level explanatory information is needed to guide character refinement. Therefore, the non-significant difference between Hidden-CoT and full UVE-PCoT does not undermine the utility of the proposed framework. Rather, it indicates that full UVE-PCoT maintains comparable predictive accuracy while offering substantially stronger interpretability and explanatory value.

We further assessed the practical magnitude of each ablation effect using Cohen's *d* for paired samples, calculated as the mean paired difference in absolute prediction error divided by the standard deviation of the paired differences. The goal of this calculation is to quantify the practical effect size of removing or altering each component, beyond mere statistical significance. Specifically, removing the overall human-likeness stage produced a large effect ($d = 0.905$), removing local-feature realism yielded a medium-to-large effect ($d = 0.724$), and removing the perceptual conflict stage produced a medium effect ($d = 0.617$). The shuffled-order variant showed a small but reliable effect ($d = 0.261$). Replacing the perceptual-conflict-stage rationale with a corrupted or generic rationale led to very large effects ($d = 1.042$ and $d = 0.957$, respectively). In contrast, the Hidden-CoT variant, which receives the same staged supervision but outputs only the final affinity score, produced a negligible and non-significant difference from the full model ($d = 0.090$, $p = 0.22$). These effect sizes indicate that the statistically significant improvements observed

for the core components are also practically meaningful, while the comparable performance of Hidden-CoT supports the use of the full UVE-PCoT setting for its interpretability without sacrificing predictive accuracy.

Taken together, the ablation results reveal that the predictive advantage of UVE-PCoT depends on both the staged decomposition and the quality of the structured rationales. The significant performance drops and the corresponding effect sizes further indicate that these components contribute not only statistically, but also practically, to affinity prediction. More specifically, overall human-likeness provides the global basis for evaluation, local-feature realism supplies the fine-grained visual evidence, and perceptual conflict detection provides a structured cue-integration step for examining whether local-feature realism jointly supports a coherent overall human-like impression before final affinity evaluation. More importantly, the control results show that UVE-PCoT is effective not merely because it adds intermediate text, but because it organizes image-grounded rationales in a meaningful perceptual structure. The comparable performance of the Hidden-CoT variant further suggests that the staged supervision itself contributes to prediction, while the visible UVE-PCoT output is retained mainly to provide interpretable cue-level explanations. These findings support the internal validity of the staged rationale design within the present dataset, but they do not by themselves establish cross-dataset generalization. This supports the role of UVE-PCoT as a structured rationale supervision framework for interpretable UVE evaluation.

5.4. Experiment 3: Evaluating Cue Identification and Interpretability in UVE-PCoT

Accurate affinity prediction alone is not sufficient to demonstrate that a model can provide interpretable evidence for the visual basis of users' affinity judgments in UVE evaluation. For design-oriented research, a more important question is whether the model can identify the visual cues associated with affinity judgments and explain how these cues are related to perceived eeriness, strangeness, discomfort, or reduced affinity. We therefore evaluate the cue identification ability of the fine-tuned UVE-PCoT model (Qwen3-VL-4B) against human reference annotations and representative general-purpose baselines.

To obtain a reliable human reference, we re-invited 30 participants from the previous study [42], including 15 male and 15 female participants, with ages ranging from 18 to 48. It should be noted that in the dataset from Li et al., which was used in Experiment 1, each participant evaluated only 25 images from a total set of 1,000. Consequently, possible prior exposure to the same humanoid character images was limited. All participants re-evaluated the test-set images by providing ratings of overall human-likeness, local-feature realism, and affinity. We then compared these newly collected ratings with the aggregated human ratings from the original dataset to assess their consistency. The results showed good agreement, with a correlation coefficient of $r = 0.82$, indicating that the new ratings are reliable for subsequent analysis.

To ensure that the stimuli cover different regions of the uncanny valley space, we adopted a 3×3 within-subject design. Based on the newly collected human ratings, test samples were divided into low, medium, and high intervals along both overall human-likeness and affinity. Specifically, overall human-likeness was divided into low (1.0–2.5), medium (2.6–3.4), and high (3.5–5.0), while affinity was divided into low (1.0–2.0), medium (2.1–3.0), and high (3.1–5.0). This yielded nine regions in total, from each of which five images were randomly sampled, resulting in 45 stimuli. In this way, the experimental set covers a balanced range of humanoid characters under different combinations of human-likeness and affinity.

During the cue-description task, participants viewed the 45 images in random order through a web interface and were asked to describe, in open-ended language, one to

three visual cues that most strongly influenced their affinity judgments. Because the human responses were expressed in unconstrained natural language, we further recruited three computer vision experts to convert these descriptions into a unified structured representation. The three experts responsible for quadruple coding in Experiment 3 were independent from the five experts involved in UVE-R validation and did not participate in the construction or verification of the UVE-R training annotations. This separation was adopted to reduce potential familiarity with the UVE-PCoT annotation style and to avoid favoring the proposed model during cue-level coding. The same coding procedure was also applied to model-generated cue descriptions, so that both human responses and model outputs could be evaluated within a common representational framework. The quadruple representation was used as a source-independent post hoc coding scheme for evaluation, rather than as a native output format specific to UVE-PCoT. Under this coding scheme, each perceptual cue was encoded as a quadruple:

$$S = (E, A, D, P), \quad (10)$$

where E denotes the facial entity, A the perceptual attribute, D the specific deviation or descriptive characteristic, and P the evaluative polarity. A detailed expert coding taxonomy is provided in Table A3. To ensure coding consistency, all responses were first coded independently by the experts and then reconciled through discussion. In cases where the original feedback was ambiguous or underspecified, the experts determined the final quadruple jointly according to the coding taxonomy and the semantic context of the response.

To reduce subjective interpretation and potential favorable coding of model outputs, the encoding process followed a rule-based and blind procedure. Before coding, all descriptions were anonymized and mixed across sources. For human participants, the experts coded the open-ended visual cue descriptions provided by the participants. For general-purpose VLM baselines, the experts coded the natural-language descriptions of visual cues related to affinity judgment. For UVE-PCoT, the coding was not performed on the complete four-stage formatted response; instead, only the natural-language rationales related to visual cues in the final affinity evaluation section were retained for coding. Model identity, output source, stage headings, and intermediate numerical scores were removed, so that the experts focused on the semantic content of the natural-language cue descriptions rather than on the original output format. The experts then used the predefined taxonomy in Table A3 as the mapping reference and coding guide for source-independent post hoc annotation, rather than as an answer table for directly matching UVE-PCoT outputs. Thus, the experts did not freely reinterpret the descriptions, but converted them according to predefined coding rules. Each of the three experts then independently encoded the descriptions without knowing whether a given text came from a human participant, UVE-PCoT, or a general-purpose VLM output. Inter-encoder consistency was assessed prior to discussion, achieving a Fleiss' Kappa of $\kappa = 0.81$, indicating strong agreement and reliable coding. This blind setting was used to avoid favoring UVE-PCoT outputs during the quadruple conversion process. After independent coding, the three encoded results were compared. Cases with identical or semantically equivalent mappings were directly retained, whereas inconsistent cases were discussed jointly. During discussion, the experts referred back to the coding taxonomy, the original image, and the original textual description when necessary to resolve ambiguity in entity selection, attribute assignment, descriptive cue wording, or polarity judgment. The final quadruple was determined by consensus after this reconciliation step.

Let $\mathcal{H}$ denote the set of human-coded quadruples and $\mathcal{M}$ the corresponding set generated by a model. We evaluate the alignment between human judgments and model

outputs at three levels. At each level, matching performance is measured by Precision, Recall, and F1-score [56]. Because each image may contain a different number of human-identified cues, we report image-wise averaged Precision, Recall, and F1-score. For the j -th image, let TP_j , FP_j , and FN_j denote the number of matched cue units, unmatched model-generated cue units, and missed human-reference cue units, respectively. Image-level Precision, Recall, and F1-score are computed as

$$P_j = \frac{TP_j}{TP_j + FP_j}, \quad R_j = \frac{TP_j}{TP_j + FN_j}, \quad F1_j = \frac{2P_jR_j}{P_j + R_j}. \quad (11)$$

The final reported values are obtained by averaging the image-level scores across all evaluated images:

$$P = \frac{1}{N} \sum_{j=1}^N P_j, \quad R = \frac{1}{N} \sum_{j=1}^N R_j, \quad F1 = \frac{1}{N} \sum_{j=1}^N F1_j. \quad (12)$$

First, at the attribute level, a match requires agreement on the entity and attribute pair (E, A) . This level evaluates whether the model identifies the same perceptual targets as the human reference. Second, at the deviation level, a match further requires agreement on the descriptive component D , yielding the triplet (E, A, D) . This level evaluates whether the model captures the same specific deviations or descriptive characteristics encoded in the human annotations. Third, at the polarity level, a correct match requires agreement on both the triplet (E, A, D) and the evaluative polarity P . This level evaluates whether the model assigns the same evaluative polarity to the identified cue as the human reference. Through this hierarchical protocol, the analysis assesses not only whether the model selects the same facial targets, but also whether it characterizes the same cue-level deviations and assigns the same evaluative polarity to them.

Tables 8–10 report the comparison results at the attribute, deviation, and polarity levels. Overall, UVE-PCoT achieves the strongest performance at all three levels, indicating a substantially higher degree of alignment with the image-wise human cue reference in both cue identification and cue interpretation.

Table 8. Comparison of cue identification performance at the attribute level.

Model	Precision ↑	Recall ↑	F1 ↑
GPT-5	0.4194	0.2737	0.3312
InternVL3.5	0.6885	0.4421	0.5385
Qwen3.5-Plus	0.5603	0.6842	0.6161
Gemini-3.1-Pro	0.6273	0.7263	0.6732
Kimi-K2.5	0.5688	0.6526	0.6078
UVE-PCoT	0.7217	0.8737	0.7905

Table 9. Comparison of cue identification performance at the deviation level.

Model	Precision ↑	Recall ↑	F1 ↑
GPT-5	0.2485	0.1824	0.2104
InternVL3.5	0.5217	0.3648	0.4294
Qwen3.5-Plus	0.4638	0.5712	0.5120
Gemini-3.1-Pro	0.5486	0.6415	0.5914
Kimi-K2.5	0.4214	0.5337	0.4709
UVE-PCoT	0.6482	0.8046	0.7180

Table 10. Comparison of cue identification performance at the polarity level.

Model	Precision ↑	Recall ↑	F1 ↑
GPT-5	0.2071	0.2113	0.1557
InternVL3.5	0.2894	0.2267	0.1636
Qwen3.5-Plus	0.2337	0.2466	0.2332
Gemini-3.1-Pro	0.2901	0.3923	0.2764
Kimi-K2.5	0.1250	0.2479	0.1584
UVE-PCoT	0.8812	0.9163	0.8914

At the attribute level, UVE-PCoT attains a Precision of 0.7217, a Recall of 0.8737, and an F1-score of 0.7905. These results indicate that the model identifies perceptual targets that are highly consistent with the human reference. In particular, the high Recall suggests that UVE-PCoT captures a broad range of the facial entities and attributes that humans consider relevant to affinity judgment. By comparison, although several general-purpose VLMs identify salient cues in some cases, their overall F1-scores remain substantially lower, indicating weaker coverage of the perceptual targets reflected in the human annotations.

At the deviation level, UVE-PCoT again yields the best result, with a Precision of 0.6482, a Recall of 0.8046, and an F1-score of 0.7180. This result shows that the model is not limited to identifying the same perceptual targets as humans, but can also capture the specific deviations or descriptive characteristics associated with those targets. This distinction is particularly important in UVE analysis, because identifying the eyes, skin, or mouth alone is insufficient unless the model can also characterize the relevant deviation, such as hollowness, waxiness, or stiffness. The consistent improvement over the general-purpose baselines suggests that structured rationale supervision helps UVE-PCoT produce cue descriptions that are more consistent with the human-coded deviations associated with affinity judgments.

At the polarity level, UVE-PCoT achieves an F1-score of 0.8914, substantially outperforming all general-purpose baselines. This result indicates that, after identifying the relevant cue and its specific deviation, the model is much more effective at assigning an evaluative polarity consistent with the human reference. In other words, UVE-PCoT not only identifies which cue is relevant, but also more accurately captures whether that cue contributes positively, neutrally, or negatively to affinity. This high polarity-level score should be interpreted as strong alignment with the structured human cue-coding reference under the proposed evaluation protocol, rather than as direct evidence that the model reproduces the internal reasoning process of human observers. By contrast, the weaker polarity-level performance of the general-purpose models suggests that they often remain at the level of feature recognition and are less reliable in mapping visual cues to their evaluative polarity in affinity judgments.

Among the general-purpose baselines, Gemini-3.1-Pro achieves the strongest overall performance, but still remains clearly below UVE-PCoT at all three levels. The largest gap appears at the polarity level, suggesting that the advantage of UVE-PCoT lies not only in cue recognition, but also in its ability to assign evaluative polarity to identified cues in a way that is consistent with human-coded references. This distinction is especially important for design-oriented applications, where a useful assistive model should not only indicate which aspects of a character appear associated with lower affinity, but also clarify how those aspects are reflected in affinity judgments.

Taken together, these results show that UVE-PCoT not only outperforms general VLMs in affinity prediction, but also aligns more closely with human reference annotations in cue identification, deviation characterization, and polarity assignment. Its consistent advantage across the attribute, deviation, and polarity levels suggests that the model is not relying

solely on superficial associations between images and scores, but benefits from structured, image-grounded rationale supervision that organizes cue-level explanations around overall human-likeness, local realism, perceptual conflict detection, and affinity evaluation.

Figure 7 presents two representative comparisons among human reference annotations, InternVL3.5 as a representative general-purpose baseline, and the proposed UVE-PCoT. In Example 1, the human-rated affinity score was 1.5, whereas InternVL3.5 substantially overestimated it as 4.0 and described the character as highly realistic and friendly, indicating that the baseline model did not adequately reflect the visual cues associated with low human affinity. By contrast, UVE-PCoT predicted an affinity score of 1.5 and showed much closer alignment with the human reference at the dimension level, particularly by identifying very low facial feature realism (1.8) and relatively limited texture realism (3.1). These predictions are consistent with the main low-affinity-related cues reflected in the human annotations, although the model still did not explicitly localize all subtle cue-level problems, such as the unnatural blue pupils and awkward mouth shape.

Example 1	Answer from InternVL3.5	Answer from UVE-PCoT	Clues from Humans
	✗ Affinity: 4 Reasoning: The character appears highly realistic, with natural-looking facial features, skin texture, and hair details. It seems to have moved beyond the lowest point of the uncanny valley, as there are no obvious unnatural elements that would provoke discomfort or aversion. Although some very subtle CGI traces may still remain, such as slightly over-smoothed skin or overly idealized lighting effects, the overall impression is friendly, acceptable, and close to that of a real human. A score of 4 indicates that the character is very close to human realism, with relatively high affinity, although it has not yet reached a fully natural or perfectly convincing level.	✓ <Overall Human-Likeness> 4.0, the character presents relatively high overall human-likeness due to its generally human-like global appearance, realistic hair, and plausible facial structure. However, this impression is weakened by the lifeless eyes and slightly unnatural skin. <Facial Features> 1.8, facial features exhibit severe unnaturalness, primarily driven by the eyes. The eyes appear vacant, lifeless, and emotionally empty, which substantially undermines the realism of the face. <Texture> 3.1, skin texture shows moderate realism but lacks natural luminosity. The surface appears somewhat dull and flat, without the subtle reflectance typically associated with living skin. <Hair> 4.2, hair exhibits high realism, with convincing layering, natural flow, and varied gloss that closely resemble real human hair. <Perceptual Conflict> The combination of highly realistic hair, moderately realistic skin, and severely lifeless eyes creates a strong cross-cue inconsistency. Because the character otherwise appears fairly human-like, the dead and vacant eyes become especially salient, producing acute perceptual conflict. <Affinity> 1.5, at a relatively high level of overall human-likeness, the lifeless eyes and deadened facial expression create a strong mismatch with the otherwise realistic appearance. This perceptual conflict substantially reduces affinity and makes the character unsettling.	Scores: Overall Human-Likeness Score: 4.0 Facial Feature Human-Likeness Score: 2.0 Texture Human-Likeness Score: 2.8 Hair Human-Likeness Score: 3.8 Affinity Score: 1.5 Cues: Eye: unnatural blue pupil Mouth: unnatural/awkward shape Expression: stiff/frozen/lifeless Skin: plastic/makeup
	✗ Affinity: 5 Reasoning: This character appears highly realistic and is almost indistinguishable from a real human. Its facial features, skin texture, eyes, lips, and other fine details all look remarkably natural and convincing. There are no obvious CGI traces or unnatural elements that would disrupt the viewer's perception. It has clearly moved beyond the uncanny valley and reached the highest level of affinity. A score of 5 indicates that the character is perceived as completely natural, friendly, and free from any features that might evoke discomfort or unease.	✓ <Overall Human-Likeness> 3.5, the character presents moderately high overall human-likeness due to its broadly human-like facial structure and natural proportions, although this impression is weakened by clearly artificial skin and hair textures. <Facial Feature> 3.3, facial features appear moderately natural overall, but the eyes are slightly oversized and overly glossy, introducing subtle digital qualities that reduce facial realism. </Facial Feature> <Texture> 2.4, the skin lacks organic complexity, showing overly uniform smoothness rather than natural micro-variation, which creates an artificial impression. <Hair> 2.4, the hair appears flat and overly uniform, lacking the layering and depth variation characteristic of real human hair, and therefore seems digitally rendered rather than organic. <Perceptual Conflicts> The character combines relatively natural facial proportions with clearly artificial skin and hair textures, while the oversized glossy eyes further amplify the mismatch. Because these conflicting cues appear within an otherwise human-like face, they generate noticeable perceptual tension. <Affinity> 2.1, At a moderately high level of overall human-likeness, the oversized glossy eyes, overly smooth skin, and uniform hair textures create a strong mismatch with the otherwise human-like face. This visual discord reduces perceived warmth and credibility, leading to low affinity.	Scores: Overall Human-Likeness Score: 3.9 Facial Feature Human-Likeness Score: 3.4 Texture Human-Likeness Score: 2.6 Hair Human-Likeness Score: 2.8 Affinity Score: 2.3 Cues: Geometry: too symmetrical Eyes: heavy/overly perfect makeup Skin: artificial/overly smooth Eyebrows: overly perfect/artificial

Figure 7. Example comparison of cue localization abilities in UVE detection by models and human annotators. Check marks indicate correctly identified cues, while wrong marks indicate missed or incorrectly identified cues.

In Example 2, the human-rated affinity score was 2.3, while InternVL3.5 again overestimated it as 5.0 and described the character as almost indistinguishable from a real human. UVE-PCoT, by contrast, produced a much closer prediction of 2.1, and its dimensional assessments were also more consistent with the human ratings, including facial feature realism (3.34 vs. 3.4), texture realism (2.4 vs. 2.6), and hair realism (2.4 vs. 2.8). Its explanation further captured the main cue-level inconsistency between relatively natural facial structure and clearly artificial skin and hair appearance. However, it still missed certain fine-grained cues highlighted in the human annotations, such as overly perfect eyebrows.

Overall, these examples provide a qualitative illustration of the quantitative results in Experiment 3. Compared with the representative general-purpose baseline, UVE-PCoT shows closer agreement with human reference ratings and cue annotations, particularly in affinity prediction and dimension-level cue assessment. At the same time, the examples also reveal that precise localization of subtle facial triggers remains challenging. Therefore, these qualitative comparisons should be interpreted as evidence of stronger cue-level alignment under the proposed evaluation protocol, rather than as proof that the model fully

reproduces human internal reasoning. This limitation suggests the need for finer-grained supervision and more detailed local-region annotations in future work.

6. Conclusions

This paper aims to advance UVE evaluation for XR humanoid characters from direct affinity-score prediction toward interpretable cue-level assessment. Specifically, it focuses on identifying which visual cues in humanoid characters are associated with users' affinity judgments from a perceptual-conflict-oriented perspective. Existing computational approaches usually predict affinity scores directly from humanoid character images, but provide limited explanation of which visual features are related to reduced affinity. UVE-PCoT addresses this limitation by organizing cue-level explanations through a perceptual-conflict-oriented structure, including overall human-likeness assessment, local-feature realism assessment, perceptual conflict detection, and final affinity evaluation. In this structure, perceptual conflict detection serves as the key step that relates local-feature realism to the overall human-like impression and the final affinity judgment. In this way, the framework provides more informative cue-level evidence than a single predicted score alone. However, this study does not empirically validate real-world design iteration or interactive decision-making with human designers. Therefore, the design implication of UVE-PCoT should be understood as providing cue-level explanatory information for preliminary assessment, rather than as direct evidence of its effectiveness in practical design workflows.

The reason for adopting a perceptual-conflict-oriented perspective is that the present study aims to make UVE evaluation interpretable at the level of visual cues. Perceptual conflict is especially relevant to this goal because it explains UVE-related responses in terms of the relationship between local visual cues and the overall human-like impression. From this perspective, the key issue is not only whether a character appears human-like as a whole, but also whether the realism of local features coherently supports that overall impression and how cue-level inconsistency may be reflected in affinity judgments. Therefore, this perspective provides a suitable basis for connecting local appearance analysis with final affinity evaluation. In this study, perceptual conflict is not treated as a directly observable psychological variable. Instead, it is operationalized at two levels. The residual-based PC index in Section 3 serves as a rating-derived operational proxy for examining whether cue-level inconsistency is associated with affinity judgments in the human-rating data. The perceptual conflict stage in UVE-PCoT further transforms this perceptual-conflict-oriented perspective into a computable assessment step, in which the model is supervised to examine whether local realism cues jointly support a coherent overall human-like impression. The goal is not to claim that the model reconstructs human internal psychological reasoning, but to provide an interpretable computational framework that combines affinity prediction with cue-level explanations.

The experimental results within the current held-out test setting consistently support the effectiveness of this framework for interpretable UVE evaluation. First, UVE-PCoT outperforms general-purpose vision-language models in affinity prediction, showing that structured perceptual decomposition and task-specific rationale supervision can improve prediction accuracy. Second, the ablation study shows that the staged structure is not merely a formal decomposition, but a functionally meaningful organization of intermediate evaluation components. In this structure, overall human-likeness provides the global reference for evaluation, local-feature realism supplies fine-grained visual evidence, and perceptual conflict detection provides a structured cue-integration step for examining whether local realism cues jointly support a coherent overall impression before final affinity evaluation. Third, UVE-PCoT achieves substantially stronger alignment with human

reference annotations in cue identification and interpretation, indicating that its advantage lies not only in predicting more accurate scores, but also in providing cue-level explanatory evidence that is more consistent with human-coded cue references under the proposed evaluation protocol.

Several limitations of the present work should be acknowledged. First, the dataset used in this study remains limited in both scale and diversity. UVE-R is constructed on top of a moderate-sized image set, and all participants come from a single cultural background. Therefore, the cross-cultural generalizability of the trained model remains untested, especially because UVE perception may vary with cultural experience, aesthetic norms, and familiarity with humanoid characters. In addition, although UVE-R provides structured rationale supervision, high-quality human annotations for subtle visual cues associated with cue-level inconsistency or reduced affinity remain limited. The current annotation scheme is based on predefined perceptual dimensions, including overall human-likeness, local-feature realism, and affinity, which may not fully capture all aspects of UVE-related perceptual responses.

The present study also has limitations related to validation scope and experimental design. The experiments focus on static facial images and do not incorporate dynamic or multimodal factors such as motion, behavior, facial expressions, or voice, all of which may influence UVE perception in real-world XR scenarios. Extending the perceptual-conflict-oriented evaluation framework to dynamic characters would require accounting for temporal and cross-modal inconsistencies, as well as interactions among appearance, expression, motion, behavior, and voice. Moreover, although each participant evaluated only a small subset of images and the presentation order was randomized, all participants had prior experience with similar experimental tasks. This prior exposure may have introduced expectation effects or response strategies beyond simple memory bias. Finally, because the present study relies on a single dataset, we cannot fully disentangle data-distribution adaptation from the model's acquisition of more generalizable cue-level evaluation ability. The ablation results, particularly the conflict-corrupted and generic-rationale variants, indicate that the performance advantage depends on semantically correct, image-grounded rationale content rather than on score-distribution memorization or staged formatting alone. However, cross-dataset validation would provide stronger evidence that the learned representations generalize beyond the current stimulus set.

Another limitation concerns the supervision of the perceptual conflict stage. The current UVE-PCoT framework does not directly use the residual-based PC index as the Stage 3 training signal. This is because the PC index is a rating-derived operational proxy rather than an independently annotated perceptual conflict label, and it is computed from overall human-likeness and local-feature realism ratings that partly overlap with the information already represented in Stage 1 and Stage 2 supervision. Instead, Stage 3 uses structured, image-grounded language rationales to provide semantic supervision for cue-level inconsistency analysis from a perceptual-conflict-oriented perspective. Therefore, the supervised formulation employed in this study should not be interpreted as evidence that the model replicates the unconscious psychological processes underlying human perception. UVE-PCoT is designed to map visual cues to structured cue-level explanations and affinity judgments using pre-existing human ratings as empirical anchors, rather than to model temporal perception, unconscious conflict detection, or individual cognitive dynamics.

Future work should address these limitations by expanding the dataset along multiple dimensions, including more diverse humanoid character types, participant populations, cultural backgrounds, and stimulus forms. In particular, future datasets should extend beyond static facial images to include dynamic and multimodal humanoid characters, so that temporal and cross-modal inconsistencies among appearance, expression, motion,

behavior, and voice can be systematically examined. Cross-dataset and cross-cultural validation will also be necessary to test whether the cue-level explanatory patterns learned by UVE-PCoT generalize beyond the current stimulus set and rating population. More fine-grained cue-level annotations are also needed to better characterize subtle visual triggers associated with affinity decline. In addition, future studies may explore PC-conditioned or joint numerical-linguistic supervision, while continuing to treat the PC index as an operational proxy rather than as a direct measure of an internal psychological state. Response-time measures, eye-tracking, and other behavioral or physiological indicators may also provide complementary evidence about observable response patterns, such as attention allocation, processing difficulty, or response hesitation, when local visual cues are inconsistent. However, these indicators should not be interpreted as direct objective measurements of perceptual conflict as an independent psychological entity. Rather, they may help examine whether the proposed cue-level conflict structure aligns with external manifestations of human rating behavior, while the present framework should still be understood as an interpretable computational tool for UVE evaluation rather than as a simulation of the full human perceptual process.

Beyond evaluation, another important future direction is to explore how UVE-PCoT could be integrated into generative AI-assisted design workflows as an evaluator module in a closed-loop process of evaluation, edit planning, generation, and re-evaluation. Specifically, UVE-PCoT could first assess an input humanoid character in terms of overall human-likeness, local-feature realism, perceptual conflict, and affinity, and then identify which local visual cues may be associated with reduced affinity. Unlike a general image-quality evaluator, UVE-PCoT would not simply suggest making all local features more realistic. Instead, it could analyze whether the realism of local features is coherent with the overall human-like impression. Its cue-level explanations could therefore be converted into editing targets, editing directions, and perceptual-consistency constraints. For example, when a character already appears highly human-like as a whole, the framework may suggest improving the naturalness of local regions such as the eyes, skin texture, or hair to reduce inconsistency caused by insufficient local realism. In contrast, when a character is more stylized or cartoon-like, the framework may suggest avoiding excessive realism in certain local features in order to maintain stylistic coherence. These cue-level explanations could then be translated into localized editing instructions for generative editing models, such as Stable Diffusion inpainting, ControlNet, or other diffusion-based local editing frameworks, to produce multiple candidate revisions. The revised images could be evaluated again by UVE-PCoT and further reviewed by designers, forming an iterative optimization loop. In this way, the potential value of UVE-PCoT lies not only in post hoc evaluation, but also in providing perceptual consistency-oriented cues for generative character refinement. However, this direction remains a future research possibility rather than a contribution demonstrated in the present experiments, and it would require validation through controlled design-iteration studies, the involvement of human designers, and user evaluations of whether such revisions improve affinity judgments.

Author Contributions: Conceptualization, X.L. and Y.X.; methodology, X.L., Y.X. and J.Q.; validation, X.L., Y.X., J.Q. and Y.Z.; formal analysis, X.L.; investigation, X.L.; resources, Y.X., J.Q., Y.Z. and C.-S.L.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, Y.X., J.Q., Y.Z. and C.-S.L.; visualization, X.L.; supervision, Y.X. and C.-S.L.; project administration, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Hunan Provincial Natural Science Foundation of China (Grant No. 2025JJ50400).

Data Availability Statement: The data used in this study are available from the corresponding author upon reasonable request.

Acknowledgments: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. During the preparation of this manuscript, the authors used GPT-5.5 Thinking for language editing and expression refinement. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Training Configuration and Reproducibility Details

To improve the reproducibility of the proposed UVE-PCoT framework, we provide a detailed summary of the training and inference configurations used for all evaluated backbone models. Tables A1 and A2 report both the shared settings applied consistently across different backbones and the model-specific settings adopted to accommodate architectural and preprocessing differences. The reported items include the LoRA configuration, frozen and trainable modules, optimizer settings, learning rate schedule, batch size, training epochs, image resolution, maximum sequence length, random seed, hardware environment, and decoding strategy.

Table A1. Training configuration and inference settings for Qwen3-VL-4B.

Configuration Item	Qwen3-VL-4B
Backbone model	Qwen3-VL-4B
LoRA rank/alpha/dropout	$r = 32$, $\alpha = 64$, dropout = 0.05
Frozen modules	The vision encoder is strictly frozen.
Trainable modules	LoRA adapters are trained while the frozen backbone parameters remain unchanged. LoRA adapters are also applied to the linear projection layer connecting the vision encoder to the language model.
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj.
Optimizer	AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay = 0.1.
Learning rate	Cosine learning-rate schedule decaying from 1×10^{-5} to 5×10^{-9} , with a warm-up ratio of 0.03.
Training epochs	10 epochs.
Effective batch size	16, with a per-device batch size of 1 and 4 gradient accumulation steps.
Image resolution	Dynamic resolution with min_pixels=3136 and max_pixels=200704.
Maximum sequence length	8192.
Random seed	42.
Hardware	Four NVIDIA RTX 3090 GPUs with gradient checkpointing enabled.
Inference input	Image and task query only. No human annotation scores, intermediate ratings, or cue labels are provided during inference.
Decoding strategy	Deterministic greedy decoding with temperature=0.0, do_sample=False, num_beams=1, and max_new_tokens=2048 under torch.inference_mode().
Ablation control	For all ablation variants based on Qwen3-VL-4B, the experimental condition is the sole independent variable. Other training and evaluation settings, including train/test splits, random seeds, and parsing rules, are kept strictly constant.

Table A2. Training configuration and inference settings for InternVL3.5-4B and LLaVA-1.5-7B.

Configuration Item	InternVL3.5-4B	LLaVA-1.5-7B
Backbone model	InternVL3.5-4B	LLaVA-1.5-7B
LoRA rank/alph/dropout	$r = 32, \alpha = 64, \text{dropout} = 0.05$	$r = 32, \alpha = 64, \text{dropout} = 0.05$
Frozen modules	The vision encoder is strictly frozen.	The vision encoder is strictly frozen.
Trainable modules	LoRA adapters are trained while the frozen backbone parameters remain unchanged. LoRA adapters are also applied to the linear projection layer connecting the vision encoder to the language model.	LoRA adapters are trained while the frozen backbone parameters remain unchanged. LoRA adapters are also applied to the linear projection layer connecting the vision encoder to the language model.
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj.	q_proj, v_proj, and the linear projection layer.
Optimizer	AdamW with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$, and weight decay = 0.1.	AdamW with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$, and weight decay = 0.1.
Learning rate	Cosine decay from 1×10^{-5} to 5×10^{-9} , with a warm-up ratio of 0.03.	2×10^{-4} .
Training epochs	10 epochs.	3 epochs.
Effective batch size	16, with a per-device batch size of 1 and 4 gradient accumulation steps.	16, with a per-device batch size of 1 and 4 gradient accumulation steps.
Image resolution	448×448 .	336×336 .
Maximum sequence length	4096.	4096.
Random seed	42.	42.
Hardware	Four NVIDIA RTX 3090 GPUs with gradient checkpointing enabled.	Four NVIDIA RTX 3090 GPUs with gradient checkpointing enabled.
Inference input	Image and task query only. No human annotation scores, intermediate ratings, or cue labels are provided during inference.	Image and task query only. No human annotation scores, intermediate ratings, or cue labels are provided during inference.
Decoding strategy	Deterministic greedy decoding with temperature=0.0, do_sample=False, num_beams=1, and max_new_tokens=2048 under torch.inference_mode().	Deterministic greedy decoding with temperature=0.0, do_sample=False, num_beams=1, and max_new_tokens=2048 under torch.inference_mode().
Ablation control	Train/test splits, random seeds, parsing rules, and other evaluation settings are kept strictly constant.	Train/test splits, random seeds, parsing rules, and other evaluation settings are kept strictly constant.

Appendix B. Expert Coding Taxonomy Reference

Table A3 summarizes the expert coding taxonomy used in Experiment 3. It provides a reference codebook for converting both human responses and model-generated cue descriptions into the structured quadruple format (E, A, D, P) . By standardizing the coding of entities, attributes, deviations, and polarity, the taxonomy establishes a unified basis for evaluating alignment between human judgments and model outputs at the attribute, deviation, and polarity levels.

Table A3. Expert coding taxonomy reference.

Domain	Entity (E)	Attribute (A)	Descriptor (D)	Polarity Example (P)
Geometry	Face	Overall Shape	Too long, Too square, Too round, Too pointed, Asymmetric	(Face, Overall Shape, Too Long, -)
Geometry	Face	Width	Too wide, Too narrow	(Face, Width, Too Wide, -)
Geometry	Face	Length	Too long, Too short	(Face, Length, Too Long, -)
Geometry	Forehead	Height	Too high, Too low	(Forehead, Height, Too High, -)
Geometry	Forehead	Width	Too wide, Too narrow	(Forehead, Width, Too Narrow, -)
Geometry	Cheekbones	Prominence	Too prominent, Flat	(Cheekbones, Prominence, Too Prominent, -)
Geometry	Jaw	Contour	Too square, Too pointed, Blurred jawline	(Jaw, Contour, Too Square, -)
Geometry	Chin	Shape	Receding, Protruding, Too pointed, Too round	(Chin, Shape, Receding, -)
Geometry	Eyes	Size	Too large, Too small, Asymmetric	(Eyes, Size, Too Large, -)
Geometry	Eyes	Shape	Too round, Too narrow, Droopy, Upturned	(Eyes, Shape, Too Droopy, -)
Geometry	Eyes	Position	Too wide-set, Too close-set, Unnatural eyebrow distance	(Eyes, Position, Too Wide-set, -)
Geometry	Eyes	Eyelid	Unnatural monolid or double eyelid, Puffy	(Eyes, Eyelid, Unnatural, -)
Geometry	Nose	Size	Too large, Too small	(Nose, Size, Too Large, -)
Geometry	Nose	Shape	Flat bridge, Hooked, Crooked, Flared nostrils, Wide nostrils	(Nose, Shape, Crooked, -)
Geometry	Nose	Bridge Height	Too high, Too low	(Nose, Bridge Height, Too High, -)
Geometry	Mouth	Size	Too wide, Too small	(Mouth, Size, Too Wide, -)
Geometry	Mouth	Lip Shape	Too thin, Too thick, Blurred lip line	(Mouth, Lip Shape, Too Thin, -)
Geometry	Mouth	Mouth Corners	Downturned, Asymmetric	(Mouth, Mouth Corners, Downturned, -)
Geometry	Ears	Size	Too large, Too small	(Ears, Size, Too Large, -)
Geometry	Ears	Position	Too high, Too low, Asymmetric	(Ears, Position, Too Low, -)
Geometry	Head	Crown	Flat crown, Low cranial top	(Head, Crown, Flat Crown, -)
Geometry	Head	Back of Head	Flat, Protruding	(Head, Back of Head, Flat, -)
Geometry	Overall Proportion	Feature Harmony	Unbalanced proportions, Uneven three-part proportion	(Overall Proportion, Feature Harmony, Unbalanced, -)
Texture	Skin	Material	Waxy, Plastic-like, Rubbery, Silicone-like	(Skin, Material, Waxy, -)
Texture	Skin	Color	Pale, Sallow, Bluish, Too red, Uneven	(Skin, Color, Pale, -)
Texture	Skin	Smoothness	Too oily, Too dry, Over-smoothed	(Skin, Smoothness, Too Oily, -)
Texture	Skin	Defects	Unrealistic pimples, Wrinkles too deep, Wrinkles too shallow, Unnatural spots	(Skin, Defects, Unrealistic, -)
Texture	Eyeball	Material	Glassy, Sticker-like, Lusterless, Cloudy	(Eyeball, Material, Glassy, -)
Texture	Eyeball	Color	Too bright, Too dark, Unnatural	(Eyeball, Color, Unnatural, -)
Texture	Teeth	Material	Too uniform, Too white, Fluorescent, Lack of translucency	(Teeth, Material, Too White, -)
Texture	Teeth	Arrangement	Too uniform, Unnaturally misaligned	(Teeth, Arrangement, Too Uniform, -)
Texture	Lighting	Reflection	Wrong specular highlights, Metallic, Cloth-like	(Lighting, Reflection, Wrong Specular, -)
Hair/Facial Hair	Hair	Structure	Blocky, Patchy, Sparse, Too voluminous	(Hair, Structure, Patchy, -)
Hair/Facial Hair	Hair	Texture	Wiry, Plastic-like, Straw-like, Lusterless	(Hair, Texture, Wiry, -)
Hair/Facial Hair	Hair	Volume	Too thin, Too thick	(Hair, Volume, Too Thin, -)
Hair/Facial Hair	Hair	Color	Unnatural dye, Harsh gradient	(Hair, Color, Unnatural, -)
Hair/Facial Hair	Hairline	Transition	Too harsh, Too straight, No baby hairs	(Hairline, Transition, Too Harsh, -)
Hair/Facial Hair	Hairline	Shape	Too high, Too low, Unnatural M-shape	(Hairline, Shape, Too High, -)
Hair/Facial Hair	Eyebrows	Shape	Too thick, Too thin, Too light, Too dark, Asymmetric	(Eyebrows, Shape, Too Thick, -)
Hair/Facial Hair	Eyebrows	Texture	Painted-on, Too uniform strands	(Eyebrows, Texture, Painted-on, -)
Hair/Facial Hair	Eyelashes	Detail	Spider-like, Too dense, Sparse, Missing eyelashes	(Eyelashes, Detail, Unnatural, -)
Hair/Facial Hair	Beard	Texture	Pasted-on, Too uniform strands, Sparse	(Beard, Texture, Pasted-on, -)
Hair/Facial Hair	Beard	Shape	Unnatural shape, Harsh boundary	(Beard, Shape, Unnatural, -)
Abstract-Perception	Eyes	Expression	Lifeless, Glazed, Staring, Vacant, Soulless	(Eyes, Expression, Lifeless, -)
Abstract-Perception	Eyes	Liveliness	Dull, Unenergetic, Robotic	(Eyes, Liveliness, Dull, -)
Abstract-Perception	Other	Unknown-Attribute	Feedback that cannot be mapped to a specific entity or attribute	(Other, Unknown Attribute, Unknown, -)

Note: The examples above illustrate negative cases only. In actual coding, the assigned polarity should follow the emotional direction expressed in the participant feedback. For ambiguous responses (e.g., “eyes look strange”), code them as (Eyes, Unknown Attribute, Unknown, -).

Appendix C. Prompt Library for the Stages of the UVE-PCoT Annotation Pipeline

The following prompt represents the complete template utilized within the annotation pipeline for the construction of the UVE-R dataset.

Prompt C.1: Prompt for Overall Human-Likeness Assessment in VLM-Based PCoT Initialization

extendedchars

Task Description:

You are an expert AI assistant specialized in evaluating humanoid character designs, focusing on overall human-likeness. Your task is to assess the character as an integrated whole and provide a perceptual explanation of how human-like the overall face appears. You must not rely on isolated local features, but instead form a holistic judgment based on the face as a coherent perceptual entity.

Based on the pre-assigned Overall Human-Likeness Score, you are required to generate image-grounded visual evidence consistent with the assigned human rating and provide a professional diagnostic rationale. A real human face serves as the reference standard corresponding to the upper bound of the realism scale (score = 5.0), where 1 indicates extremely non-human appearance and 5 indicates appearance that is nearly indistinguishable from a real human.

Evaluation Focus:

The evaluation should emphasize the global perceptual impression of human-likeness, including the following aspects:

1. Global Facial Configuration

Assess whether the face is organized in a manner consistent with human facial structure at a coarse level, without focusing on fine-grained local accuracy.

2. Holistic Human Resemblance

Evaluate the overall resemblance to a human face as a unified percept, rather than a combination of independent features.

3. Perceptual Naturalness and Lifelikeness

Assess whether the face appears natural, lifelike, and plausible as a human entity, rather than artificial, stylized, or synthetic.

4. Category Perception

Evaluate whether the face is perceived as clearly non-human, ambiguously human-like, or close to a real human.

5. Perceptual Coherence

Assess whether different visible cues jointly support a consistent human-like impression or introduce ambiguity in category perception.

Input Format:

Image, Overall Human-Likeness Score

Output Format:

<Overall Human-Likeness>

[Provide a step-by-step perceptual analysis of the character's overall human-likeness. Explain how the face as a whole aligns with or deviates from realistic human appearance, and describe the observable visual evidence that is consistent with the assigned human rating \${Overall Human-Likeness Score}.]

Therefore, the overall human-likeness score for this character is: \${Overall Human-Likeness Score}

</Overall Human-Likeness>

Important Notes:

1. Rating-Anchor Consistency Constraint: The input score is used as a human-rating anchor. Construct a diagnostic rationale based on observable image features that is consistent with the assigned score; the score itself should not be modified or re-estimated.

2. The analysis must focus on the overall perceptual impression of the face as a whole, rather than isolated local features. Individual cues may be referenced only insofar as they contribute to the global human-likeness perception.

3. All explanations must remain image-grounded, perceptual, and consistent with the assigned human-rating anchor.

4. Strict adherence to the specified output format is required. No additional text beyond the defined structure is allowed.

5. The explanation within the <Overall Human-Likeness> tag should identify observable visual cues consistent with the human-rating anchor and form a structured diagnostic explanation.

Prompt C.2: Prompt for Facial Geometry Realism Assessment in VLM-Based PCoT Initialization

extendedchars

Task Description:

You are an expert AI assistant specialized in evaluating humanoid character designs, focusing on facial geometry and feature proportions. Your task is to exclude factors such as skin texture, lighting, and hair, and focus solely on facial proportions and the structural layout of individual features (eyes, nose, mouth, eyebrows, teeth, etc.). You must provide a detailed and objective geometric analysis. Based on the pre-assigned Facial Geometry Realism Score, you are required to generate image-grounded visual evidence consistent with the assigned human rating and provide a professional diagnostic rationale. A real human face (score = 5.0) serves as the absolute reference standard (1 indicates extremely non-human, and 5 indicates indistinguishable from a real human).

Evaluation Dimensions:

1. Global Facial Proportions

Evaluate the vertical segmentation (Three Courts: hairline--brow, brow--nose base, nose--chin) for proportional balance; Evaluate the horizontal segmentation (Five Eyes rule) to determine whether facial width aligns with standard human spacing; Assess overall facial proportion consistency and global structural balance.

2. Eyes

Assess interocular distance, which is expected to approximate one eye width; Evaluate eye size relative to face width (detect proportional exaggeration or reduction); Analyze vertical alignment with eyebrows and nose base; Check bilateral symmetry and contour consistency.

3. Nose

Evaluate nose width relative to interocular distance (expected to align with inner eye corners); Assess nose length within the middle facial third; Check alignment along the vertical facial axis; Identify geometric simplifications, exaggerations, or structural deviations.

4. Mouth and Jaw

Assess mouth width relative to pupil alignment (expected to align with iris centers); Evaluate vertical placement within the lower facial third; Analyze jaw shape and chin proportion relative to overall facial geometry; Identify disproportionate scaling or structural inconsistencies.

5. Eyebrows

Evaluate alignment with the eye contour and supraorbital ridge; Assess spacing between eyebrows and eyes; Check symmetry and curvature consistency.

6. Facial Feature Alignment

Evaluate whether key features (eyes, nose, mouth) align along a coherent vertical midline; Assess horizontal alignment consistency (e.g., both eyes positioned on the same axis); Identify structural misalignment or geometric offsets.

7. Symmetry and Structural Consistency

Assess bilateral symmetry across the vertical axis; Identify local asymmetries and evaluate whether they fall within natural human variation or indicate structural distortion; Evaluate overall structural coherence of the face.

Input Format:

Image, Facial Geometry Realism Score

Output Format:

<Facial Geometry>

[Provide a step-by-step technical analysis. Pinpoint why the geometry aligns with or deviates from a real human facial structure standard, and describe observable visual evidence that is consistent with the assigned human rating \${Facial Geometry Realism Score}.]

Therefore, the facial geometry realism score for this character is: \${Facial Geometry Realism Score}

</Facial Geometry>

Important Notes:

1. Rating-Anchor Consistency Constraint: The input score is used as a human-rating anchor. Construct a diagnostic rationale based on observable structural features that is consistent with the assigned human-rating anchor; the score itself should not be modified or re-estimated.
2. All justifications must be grounded in observable structural features. For low scores, the rationale must explicitly identify concrete geometric deviations from realistic human facial structure.
3. Strict adherence to the specified output format is required. All explanations must remain image-grounded and perceptual, consistent with the assigned human-rating anchor.
4. The explanation within the <Facial Geometry> tag should identify observable visual cues consistent with the human-rating anchor and form a structured diagnostic explanation.

Prompt C.3: Prompt for Texture Realism Assessment in VLM-Based PCoT Initialization

extendedchars

Task Description:

You are an expert AI assistant specialized in evaluating humanoid character designs, focusing on facial texture realism. Your task is to exclude factors related to facial geometry (e.g., proportions, feature placement, and structural layout), and instead focus solely on surface-level visual properties, including skin texture, material appearance, reflectance, and fine-grained details (e.g., pores, wrinkles, and shading consistency). Based on the pre-assigned Facial Texture Realism Score, you are required to generate image-grounded visual evidence consistent with the assigned human rating and provide a professional diagnostic rationale. A real human face serves as the reference standard corresponding to the upper bound of the realism scale (score = 5.0), where 1 indicates extremely artificial appearance and 5 indicates indistinguishable from real human skin and material properties.

Evaluation Dimensions:

1. Skin Microstructure

Evaluate the presence and realism of fine-scale details such as pores, wrinkles, and subtle irregularities; Assess whether the skin exhibits natural stochastic variation rather than uniform smoothness; Identify over-smoothing, plastic-like surfaces, or synthetic noise patterns.

2. Material Realism

Assess whether the skin appears biologically plausible rather than synthetic (e.g., wax, plastic, or silicone-like); Evaluate subsurface scattering effects and translucency consistency; Identify unrealistic material cues such as overly reflective or overly matte surfaces.

3. Shading and Illumination Consistency

Evaluate whether shading transitions are smooth and physically plausible across the face; Assess consistency between local shading and global light direction; Identify abrupt gradients, patchy shading, or non-physical illumination artifacts.

4. Color Distribution and Variation

Assess whether skin tone exhibits natural variation (e.g., redness around cheeks, darker tones near eyes); Evaluate color consistency across facial regions; Identify flat, overly uniform, or artificially segmented color patterns.

5. Detail Consistency and Resolution

Evaluate whether fine details (pores, lines, micro-shadows) are consistent across regions; Identify mismatches such as highly detailed areas adjacent to overly smooth regions; Detect texture tiling, repetition, or resolution inconsistencies.

6. Local Surface Quality

Assess texture realism in perceptually sensitive regions such as around the eyes, nose, and mouth; Evaluate whether these areas exhibit appropriate fine detail and smooth material transitions; Identify artifacts such as blurred boundaries, unrealistic transitions, or synthetic smoothing.

Input Format:

Image, Texture Realism Score

Output Format:

<Texture>

[Provide a step-by-step technical analysis. Explain how the facial texture aligns with or deviates from realistic human skin appearance, and describe observable visual evidence that is consistent with the assigned human rating \${Texture Realism Score}.]

Therefore, the facial texture realism score for this character is: \${Texture Realism Score}

</Texture>

Important Notes:

1. Rating-Anchor Consistency Constraint: The input score is used as a human-rating anchor. Construct a diagnostic rationale based on observable surface-level features that is consistent with the assigned human-rating anchor; the score itself should not be modified or re-estimated.

2. All justifications must be grounded in observable surface-level properties. For low scores, the rationale must explicitly identify concrete deviations from realistic human skin appearance.

3. Strict adherence to the specified output format is required. All explanations must remain image-grounded and perceptual, consistent with the assigned human-rating anchor.

4. The explanation within the <Texture> tag should identify observable visual cues consistent with the human-rating anchor and form a structured diagnostic explanation.

Prompt C.4: Prompt for Hair Realism Assessment in VLM-Based PCoT Initialization

extendedchars

Task Description:

You are an expert AI assistant specialized in evaluating humanoid character designs, focusing on hair and facial hair realism. Your task is to exclude factors related to facial geometry and general skin texture, and instead focus exclusively on hair-related visual properties, including scalp hair, eyebrows, eyelashes, and facial hair (if present). You must provide a detailed and objective analysis of hair realism. Based on the pre-assigned Hair Realism Score, you are required to generate image-grounded visual evidence consistent with the assigned human rating and provide a professional diagnostic rationale. A real human serves as the reference standard corresponding to the upper bound of the realism scale (score = 5.0), where 1 indicates extremely artificial or implausible hair appearance and 5 indicates indistinguishable from real human hair.

Evaluation Dimensions:

1. Hair Strand Structure

Evaluate whether individual hair strands are discernible or realistically implied; Assess variation in strand thickness, continuity, and structural definition; Identify artifacts such as overly smooth masses, painted textures, or synthetic strand patterns.

2. Density and Distribution

Assess whether hair density varies naturally across regions (e.g., denser in central areas and softer at boundaries); Evaluate distribution consistency across scalp hair, eyebrows, eyelashes, and facial hair when present; Identify unnatural uniformity, excessive sparsity, or overcrowded distribution.

3. Hairline and Boundary Transition

Evaluate the realism of transitions between hair and skin, such as along the scalp, forehead, and eyebrow edges; Assess whether boundaries appear soft and irregular rather than sharply cut or artificially segmented; Identify abrupt transitions, unnatural edges, or halo-like artifacts.

4. Flow and Directionality

Assess whether hair follows plausible growth directions and flow patterns; Evaluate directional coherence across the scalp and facial hair regions; Identify chaotic, inconsistent, or physically implausible strand orientation.

5. Local Hair Quality in Sensitive Regions

Assess the realism of eyebrows, eyelashes, and facial hair (if present); Evaluate whether these regions exhibit appropriate density, curvature, shape coherence, and strand-level detail; Identify simplification, over-stylization, or missing fine structures in these perceptually sensitive areas.

6. Detail Consistency and Resolution

Assess whether hair detail resolution is consistent across different regions; Identify mismatches such as highly detailed eyebrows but blurred scalp hair; Detect repetition, tiling, or local degradation in hair detail.

7. Global Hair Coherence

Evaluate whether all hair-related elements (scalp hair, eyebrows, eyelashes, and facial hair if present) form a coherent perceptual whole; Identify inconsistencies in style, density, or realism level across regions; Assess whether the overall hair appearance aligns with realistic human expectations.

Input Format:

Image, Hair Realism Score

Output Format:

<Hair>

[Provide a step-by-step technical analysis. Explain how the hair appearance aligns with or deviates from realistic human hair, and describe observable visual evidence that is consistent with the assigned human rating \${Hair Realism Score}.]

Therefore, the hair realism score for this character is: \${Hair Realism Score}

</Hair>

Important Notes:

1. Rating-Anchor Consistency Constraint: The input score is used as a human-rating anchor. Construct a diagnostic rationale based on observable hair features that is consistent with the assigned human-rating anchor; the score itself should not be modified or re-estimated.
2. All justifications must be grounded in observable hair-related features. For low scores, the rationale must explicitly identify concrete deviations from realistic human hair appearance.
3. Strict adherence to the specified output format is required. All explanations must remain image-grounded and perceptual, consistent with the assigned human-rating anchor.
4. The explanation within the <Hair> tag should identify observable visual cues consistent with the human-rating anchor and form a structured diagnostic explanation.

Prompt C.5: Prompt for Perceptual Conflict Detection in VLM-Based PCoT Initialization

extendedchars

Task Description:

You are an expert AI assistant specialized in identifying perceptual conflict in humanoid character designs. Your task is to analyze whether the realism of local facial features is consistent with the overall human-likeness impression, and whether different local features remain perceptually coordinated with one another. Based on the provided summaries of overall human-likeness and local-feature realism (facial geometry, texture, and hair), you are required to infer and explain the presence of perceptual conflict. Perceptual conflict refers to inconsistencies that weaken the overall perceptual coherence of a humanoid face. Such conflict arises when the overall human-likeness establishes a certain expectation of realism, but one or more local features deviate from that expectation, or when different local features imply mismatched levels of realism. The analysis must follow the staged perceptual pathway: overall human-likeness → local-feature realism → perceptual conflict. This stage should infer conflict primarily from the relationship among the provided stage summaries and scores, while remaining grounded in the input image.

Evaluation Focus:

1. Consistency with Overall Human-Likeness

Evaluate whether facial geometry, texture, and hair realism are individually consistent with the overall human-likeness impression.

2. Cross-Feature Consistency

Assess whether geometry, texture, and hair exhibit coordinated levels of realism or imply conflicting realism cues.

3. Local Deviations from Global Expectation

Identify which local features deviate from the realism level implied by the overall human-likeness.

4. Conflict Localization

Identify specific cues or regions that contribute to perceptual conflict (e.g., highly realistic structure with artificial texture, or realistic skin with implausible hair).

5. Perceptual Coherence

Assess whether the face appears perceptually integrated as a coherent human-like entity or fragmented due to conflicting cues.

Input Format:

Image, Overall Human-Likeness Summary, Facial Geometry Realism Summary, Texture Realism Summary, Hair Realism Summary, Overall Human-Likeness Score, Facial Geometry Realism Score, Texture Realism Score, Hair Realism Score

Output Format:

<Perceptual Conflict>

[Provide a structured analysis of whether perceptual conflict exists. Explain how local-feature realism relates to the overall human-likeness impression, identify inconsistencies between geometry, texture, and hair, and describe which cues weaken the overall perceptual coherence of the face.]

</Perceptual Conflict>

Important Notes:

1. This task does not involve predicting or justifying a numerical score.

2. Do not introduce new visual evidence that is not supported by the image or the provided summaries. All explanations must be grounded in the given inputs.

3. The analysis must focus on inconsistencies between overall human-likeness and local-feature realism, as well as inconsistencies among local features.

4. The analysis must strictly follow the staged perceptual pathway: overall human-likeness → local-feature realism → perceptual conflict.

5. Do not directly predict or infer affinity in this section.

6. Strict adherence to the specified output format is required. Output only the content inside the <Perceptual Conflict> tag, with no additional text.

Prompt C.6: Prompt for Affinity Evaluation in VLM-Based PCoT Initialization

extendedchars

Task Description:

You are an expert AI assistant specialized in evaluating affinity responses to humanoid character designs. Your task is to generate a structured diagnostic explanation of the assigned Affinity Score by integrating the preceding stages: overall human-likeness, local-feature realism, and perceptual conflict. Based on the provided summaries and scores, you should explain how the overall human-likeness establishes an initial affinity expectation, how facial geometry, texture, and hair realism reinforce or weaken this expectation, and how perceptual conflict modulates the final affinity response. The Affinity Score should be used as a human-rating anchor, and the explanation should remain grounded in the input image and consistent with the outputs of the previous stages.

Evaluation Focus:

1. Baseline from Overall Human-Likeness

Explain how the overall human-likeness establishes an initial expectation of affinity.

2. Influence of Local-Feature Realism

Assess whether facial geometry, texture, and hair realism reinforce or weaken this baseline expectation.

3. Role of Perceptual Conflict

Analyze whether perceptual conflict exists and how it suppresses or alters the expected affinity.

4. Perceptual Integration

Evaluate whether the face is perceived as coherent and natural or disrupted and inconsistent.

5. Final Affinity Formation

Explain how the interaction between overall human-likeness, local-feature realism, and perceptual conflict leads to the final affinity score.

Input Format:

Image, Overall Human-Likeness Summary, Facial Geometry Realism Summary, Texture Realism Summary, Hair Realism Summary, Overall Human-Likeness Score, Facial Geometry Realism Score, Texture Realism Score, Hair Realism Score, Affinity Score

Output Format:

<Affinity>

[Provide a structured explanation of the affinity response. Explain how overall human-likeness establishes a baseline, how local-feature realism and perceptual conflict influence this baseline, and describe observable visual and perceptual evidence consistent with the assigned human-rating anchor \${Affinity Score}.]

Therefore, the affinity score for this character is: \${Affinity Score}

</Affinity>

Important Notes:

1. Rating-Anchor Consistency Constraint: The Affinity Score is used as a human-rating anchor. Construct a diagnostic rationale based on observable evidence that is consistent with the assigned human-rating anchor; the score itself should not be modified or re-estimated.

2. The explanations must strictly follow the staged perceptual pathway:

overall human-likeness → local-feature realism → perceptual conflict → affinity.

3. Do not infer affinity directly from visual appearance without referencing the intermediate perceptual stages.

4. Perceptual conflict must be treated as a key factor that modulates the final affinity response.

5. All explanations must be consistent with the outputs of previous stages and grounded in the provided inputs.

6. Strict adherence to the specified output format is required. Output only the content inside the <Affinity> tag, with no additional text.

Prompt C.7: Prompt for LLM-Based PCoT Fusion

extendedchars

Task Description:

You are an expert AI assistant for reviewing and fusing perceptual reasoning chains in humanoid character evaluation. Your task is to examine two candidate perceptual chains of thought, named Kimi_CoT and Qwen_CoT, that were generated for the same image and the same human annotation scores. You must treat the image as visual evidence and the human annotation scores as rating anchors for consistency. The two candidate reasoning chains are not visual evidence; they are candidate perceptual interpretations that must be verified against the image and the provided scores.

Your task is to:

1. Compare Kimi_CoT and Qwen_CoT section by section.
2. Detect conflicts, ambiguities, hallucinated visual cues, or modality-inconsistent descriptions.
3. Identify which perceptual dimensions (e.g., overall human-likeness, geometry, texture, hair, perceptual conflict, affinity) lead to divergent interpretations.
4. Evaluate whether each candidate reasoning chain follows the intended staged perceptual pathway: overall human-likeness → local-feature realism (facial geometry, texture, hair) → perceptual conflict → affinity
5. Check whether the reasoning is logically consistent, complete, and traceable.
6. If conflicts or issues exist, explain them and suggest improvements.
7. Produce one final fused perceptual reasoning chain that is clear, coherent, and aligned with the image and the human annotation scores.
8. Ensure that all supervised sections in the fused output remain consistent with their corresponding human-rating anchors.

Input Format:

Image, Human Annotation Scores, Kimi_CoT, Qwen_CoT

Output Format:

1. CoT conflict: 1 if correctly detected, 0 if consistent.
2. Conflict_location: Which sections have conflicts: <Overall Human-Likeness>, <Facial Geometry Realism>, <Texture Realism>, <Hair Realism>, <Perceptual Conflict>, or <Affinity>. Enter 'None' if no conflict.
3. CoT conflict reason: Point out where conflicts exist in the chain of thought. If no issues, enter 'None'.
4. Suggestions for improvement: Suggestions for improving the reasoning chain. If no issues, enter 'None'.
5. Fused_PCoT: Unified reasoning chain following original PCoT format.
6. Alignment_verification: All supervised sections MUST remain consistent with their corresponding human-rating anchors: <Overall Human-Likeness>, <Facial Geometry Realism>, <Texture Realism>, <Hair Realism> and <Affinity>.

Important Notes:

1. Use the image as visual evidence and the human annotation scores as rating anchors for consistency.
2. Do not treat any content from Kimi_CoT or Qwen_CoT as visual evidence unless it is clearly supported by the image.
3. Do not introduce new visual details that are not supported by the image.
4. Do not modify or re-estimate any human annotation score; use each score only as a human-rating anchor for constructing the fused diagnostic rationale.
5. All supervised sections (<Overall Human-Likeness>, <Facial Geometry Realism>, <Texture Realism>, <Hair Realism>, and <Affinity>) must remain consistent with their corresponding human-rating anchors.
6. The explanation in each supervised section should identify observable visual or perceptual evidence consistent with its assigned human-rating anchor.
7. If either candidate reasoning chain is inconsistent with the image, the staged perceptual pathway, or the human annotation scores, you must correct it in the fused output.
8. The fused output must be a single coherent perceptual reasoning chain, not a comparison, mixture, or concatenation of the two candidate inputs.
9. Do not copy Kimi_CoT or Qwen_CoT directly.
10. Ensure that the fused reasoning strictly follows the staged perceptual pathway: overall human-likeness → local-feature realism → perceptual conflict → affinity
11. Output must be valid JSON only, with no additional text or formatting.

References

1. Strohmman, T.; Siemon, D.; Khosrawi-Rad, B.; Robra-Bissantz, S. Toward a design theory for virtual companionship. *Hum. Comput. Interact.* **2023**, *38*, 194–234.
2. Sylaiou, S.; Fidas, C. Virtual humans in museums and cultural heritage sites. *Appl. Sci.* **2022**, *12*, 9913.
3. Fink, M.C.; Robinson, S.A.; Ertl, B. AI-based avatars are changing the way we learn and teach: Benefits and challenges. *Front. Educ.* **2024**, *9*, 1416307. <https://doi.org/10.3389/educ.2024.1416307>.

4. Reinhardt, J.; Hillen, L.; Wolf, K. Embedding conversational agents into ar: Invisible or with a realistic human body? In Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction, Sydney, Australia, 9–12 February 2020; pp. 299–310.
5. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
6. Cao, Y.; Cao, Y.P.; Han, K.; Shan, Y.; Wong, K.Y.K. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 958–968.
7. Kolotouros, N.; Alldieck, T.; Corona, E.; Bazavan, E.G.; Sminchisescu, C. Instant 3d human avatar generation using image diffusion models. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 177–195.
8. Mori, M. The uncanny valley. *Energy* **1970**, *7*, 33–35.
9. Mori, M.; MacDorman, K.F.; Kageki, N. The Uncanny Valley [From the Field]. *IEEE Robot. Autom. Mag.* **2012**, *19*, 98–100. <https://doi.org/10.1109/MRA.2012.2192811>.
10. MacDorman, K.F.; Ishiguro, H. Opening Pandora’s uncanny Box. *Interact. Stud.* **2006**, *7*, 361–368. <https://doi.org/10.1075/is.7.3.10mac>.
11. Riek, L.D.; Rabinowitch, T.C.; Chakrabarti, B.; Robinson, P. Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In Proceedings of the 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–6. <https://doi.org/10.1109/ACII.2009.5349423>.
12. Nowak, K.L.; Biocca, F. The Effect of the Agency and Anthropomorphism on Users’ Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence Teleoperators Virtual Environ.* **2003**, *12*, 481–494. <https://doi.org/10.1162/105474603322761289>.
13. Barreda-Ngeles, M.; Aleix-Guillaume, S.; Pereda-Baos, A. An “Empathy Machine” or a “Just-for-the-Fun-of-It” Machine? Effects of Immersion in Nonfiction 360-Video Stories on Empathy and Enjoyment. *Cyberpsychology, Behav. Soc. Netw.* **2020**, *23*, 683–688.
14. Grimshaw, M. The audio Uncanny Valley: Sound, fear and the horror game. In *Proceedings of the Audio Mostly*, Glasgow, UK, 2–3 September 2009; pp. 21–26.
15. Tinwell, A.; Grimshaw, M.; Nabi, D.A.; Williams, A. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Comput. Hum. Behav.* **2011**, *27*, 741–749. <https://doi.org/10.1016/j.chb.2010.10.018>.
16. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81. <https://doi.org/10.1007/s12369-008-0001-3>.
17. Saygin, A.P.; Chaminade, T.; Ishiguro, H.; Driver, J.; Frith, C. The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Soc. Cogn. Affect. Neurosci.* **2012**, *7*, 413–422. <https://doi.org/10.1093/scan/nsr025>.
18. Mustafa, M.; Guthe, S.; Tauscher, J.P.; Goesele, M.; Magnor, M. How Human Am I? EEG-based Evaluation of Virtual Characters. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, Colorado, USA, 6–11 May 2017; CHI ’17; pp. 5098–5108. <https://doi.org/10.1145/3025453.3026043>.
19. Cheetham, M.; Pavlovic, I.; Jordan, N.; Suter, P.; Jancke, L. Category Processing and the human likeness dimension of the Uncanny Valley Hypothesis: Eye-Tracking Data. *Front. Psychol.* **2013**, *4*, 108. <https://doi.org/10.3389/fpsyg.2013.00108>.
20. Burleigh, T.J.; Schoenherr, J.R. A reappraisal of the uncanny valley: Categorical perception or frequency-based sensitization? *Front. Psychol.* **2015**, *5*, 1488. <https://doi.org/10.3389/fpsyg.2014.01488>.
21. Ho, C.C.; MacDorman, K.F. Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. *Int. J. Soc. Robot.* **2017**, *9*, 129–139. <https://doi.org/10.1007/s12369-016-0380-9>.
22. Igaue, T.; Hayashi, R. Signatures of the uncanny valley effect in an artificial neural network. *Comput. Hum. Behav.* **2023**, *146*, 107811.
23. Jiang, R.; Chen, C.W. Multimodal LLMs Can Reason about Aesthetics in Zero-Shot. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025; ACM: New York, NY, USA, 2025; pp. 6634–6643. <https://doi.org/10.1145/3746027.3754961>.
24. Ke, J.; Ye, K.; Yu, J.; Wu, Y.; Milanfar, P.; Yang, F. Vila: Learning image aesthetics from user comments with vision-language pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10041–10051.
25. Moore, R.K. A Bayesian explanation of the ‘Uncanny Valley’ effect and related psychological phenomena. *Sci. Rep.* **2012**, *2*, 864.
26. MacDorman, K.F.; Chattopadhyay, D. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* **2016**, *146*, 190–205.
27. Diel, A.; Lewis, M. Rethinking the uncanny valley as a moderated linear function: Perceptual specialization increases the uncanniness of facial distortions. *Comput. Hum. Behav.* **2024**, *157*, 108254.

28. Wang, W.; Gao, Z.; Gu, L.; Pu, H.; Cui, L.; Wei, X.; Liu, Z.; Jing, L.; Ye, S.; Shao, J.; et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv* **2025**, arXiv:2508.18265. <http://arxiv.org/abs/2508.18265>.
29. Team, Q. Qwen3. 5-omni technical report. *arXiv* **2026**, arXiv:2604.15804.
30. MacDorman, K.F.; Ishiguro, H. The uncanny advantage of using androids in cognitive and social science research. *Interact. Stud. Soc. Behav. Commun. Biol. Artif. Syst.* **2006**, *7*, 297–337.
31. MacDorman, K.F.; Green, R.D.; Ho, C.C.; Koch, C.T. Too real for comfort? Uncanny responses to computer generated faces. *Comput. Hum. Behav.* **2009**, *25*, 695–710.
32. MacDorman, K.F. Mortality salience and the uncanny valley. In *Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots*, Tsukuba, Japan, 5–7 December 2005; pp. 399–405. <https://doi.org/10.1109/ICHR.2005.1573600>.
33. Yamada, Y.; Kawabe, T.; Ihaya, K. Categorization difficulty is associated with negative evaluation in the “uncanny valley” phenomenon. *Jpn. Psychol. Res.* **2013**, *55*, 20–32.
34. Seyama, J.; Nagayama, R.S. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence* **2007**, *16*, 337–351.
35. Meah, L.F.; Moore, R.K. The uncanny valley: A focus on misaligned cues. In *Social Robotics*; Beetz, M., Johnston, B., Williams, M.A., Eds.; Springer: Cham, Switzerland, 2014; pp. 256–265. https://doi.org/10.1007/978-3-319-11973-1_26.
36. Gray, H.M.; Gray, K.; Wegner, D.M. Dimensions of mind perception. *Science* **2007**, *315*, 619–619.
37. Ho, C.C.; MacDorman, K.F. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Comput. Hum. Behav.* **2010**, *26*, 1508–1518. <https://doi.org/10.1016/j.chb.2010.05.015>.
38. Mathur, M.B.; Reichling, D.B. Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition* **2016**, *146*, 22–32. <https://doi.org/10.1016/j.cognition.2015.09.008>.
39. Mathur, M.B.; Reichling, D.B.; Lunardini, F.; Geminiani, A.; Antonietti, A.; Ruijten, P.A.; Levitan, C.A.; Nave, G.; Manfredi, D.; Bessette-Symons, B.; et al. Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley. *Comput. Hum. Behav.* **2020**, *103*, 21–30. <https://doi.org/10.1016/j.chb.2019.08.029>.
40. Bartneck, C.; Kanda, T.; Ishiguro, H.; Hagita, N. Is The Uncanny Valley An Uncanny Cliff? In *Proceedings of the RO-MAN 2007—The 16th IEEE International Symposium on Robot and Human Interactive Communication*, Jeju Island, Republic of Korea, 26–29 August 2007; pp. 368–373. <https://doi.org/10.1109/ROMAN.2007.4415111>.
41. Kim, B.; Bruce, M.; Brown, L.; Visser, E.d.; Phillips, E. A Comprehensive Approach to Validating the Uncanny Valley using the Anthropomorphic RoBOT (ABOT) Database. In *Proceedings of the 2020 Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, 24 April 2020; pp. 1–6. <https://doi.org/10.1109/SIEDS49339.2020.9106675>.
42. Li, X.; Xiao, Y.; Zheng, Y.; Qiao, J.; Leung, C.S. Revisiting the Uncanny Valley Effect: A data-driven analysis with curve fitting perspective. In *Proceedings of the Design Research Society 2024*, Boston, MA, USA, 24–28 June 2024; pp. 23–28. <https://doi.org/10.21606/drs.2024.422>.
43. Dal Molin, G.P.; Nomura, F.M.; Dalmoro, B.M.; de A. Araújo, V.F.; Musse, S.R. Can we estimate the perceived comfort of virtual human faces using visual cues? In *Proceedings of the 2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, Virtual, 27–29 January 2021; pp. 366–369. <https://doi.org/10.1109/ICSC50631.2021.00085>.
44. Dal Molin, G.P.; Musse, S.R.; Feijo, B.; Braun, A. Predicting Uncanny Valley Perception in Virtual Humans Faces Through Computer Vision Techniques. *SSRN* **2024**. <http://doi.org/10.2139/ssrn.4844650>.
45. Li, K.; Po, L.M.; Yang, H.; Xu, X.; Liu, K.; Zhao, Y. AesBiasBench: Evaluating Bias and Alignment in Multimodal Language Models for Personalized Image Aesthetic Assessment. *arXiv* **2025**, arXiv:2509.11620. <http://arxiv.org/abs/2509.11620>.
46. Chen, Y.; Han, J.; Bai, T.; Tong, S.; Kokkinos, F.; Torr, P. From Pixels to Feelings: Aligning MLLMs with Human Cognitive Perception of Images. *arXiv* **2025**, arXiv:2511.22805. <http://arxiv.org/abs/2511.22805>.
47. Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; Smola, A.J. Multimodal Chain-of-Thought Reasoning in Language Models. *Trans. Mach. Learn. Res.* **2024**.
48. Jiang, Z.; Liu, J.; Li, H.; Liu, Y.; Song, Y. Multi-Stage Image Aesthetic Assessment via Chain-of-Thought Reasoning. *IEEE Access* **2025**, *13*, 131159–131165. <https://doi.org/10.1109/ACCESS.2025.3590925>.
49. Wang, S.; Hu, M.; Li, Q.; Safari, M.; Qiu, R.L.; Yang, X. Capabilities of GPT-5 on multimodal medical reasoning. In *Proceedings of the Medical Imaging 2026: Imaging Informatics*, 139300E; SPIE: Bellingham, WA, USA, 2026; Volume 13930. <https://doi.org/10.1117/12.3086794>.
50. Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv* **2025**, arXiv:2507.06261. <http://arxiv.org/abs/2507.06261>.
51. Team, K.; Bai, T.; Bai, Y.; Bao, Y.; Cai, S.; Cao, Y.; Charles, Y.; Che, H.; Chen, C.; Chen, G.; et al. Kimi K2. 5: Visual Agentic Intelligence. *arXiv* **2026**, arXiv:2602.02276.

52. Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv* **2024**, arXiv:2412.10302. <http://arxiv.org/abs/2412.10302>.
53. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685. <http://arxiv.org/abs/2106.09685>.
54. James, G. *An Introduction to Statistical Learning with Applications in R*; Springer: New York, NY, USA, 2013.
55. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82.
56. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.