

Article

Template-Based Evaluation of Stable Diffusion via Attention Maps

Haruno Fusa ^{1,*} , Chonho Lee ^{1,*} , Sakuei Onishi ¹ , Kanshin Fusa ² and Hiromitsu Shiina ¹

¹ Department of Information Science and Engineering, Okayama University of Science, Okayama 700-0005, Okayama, Japan

² Department of Communication and Information System Program, National Institute of Technology Tsuyama College, Tsuyama 708-8509, Okayama, Japan

* Correspondence: r24sdj2fu@ous.jp (H.F.); leech@ous.ac.jp (C.L.)

Abstract

Text-to-image models such as Stable Diffusion (SD) require comprehensive, fine-grained, and high-precision methods for evaluating text–image alignment. A prior method, the text–image alignment metric (TIAM), employs a template-based approach for fine-grained, high-precision evaluation; however, it is restricted to objects and colors, limiting its comprehensiveness. This study extends the TIAM by incorporating attention maps and vision–language models to deliver a fine-grained and high-precision evaluation framework that goes beyond colors and objects to include attributes, actions, and positions. In our experiments, we analyze the evaluation scores of images generated by the proposed method and compare them with human judgments. The results demonstrate that the proposed method outperforms existing methods, exhibiting a stronger correlation with human judgments ($r = 0.853$, $p < 10^{-48}$). In addition, we applied the proposed method to evaluate the generation abilities of three SD models (i.e., SD1.4, SD2, and SD3.5). Each experiment used over 900 images, totaling 9858 images across all experiments to ensure statistical significance. The results indicate that SD3.5 exhibits superior expressiveness compared with SD1.4 and SD2. Nevertheless, for more complex tasks such as multi-attribute generation or multi-action generation, limitations in text–image alignment remain evident.

Keywords: image synthesis; diffusion processes; text-to-image; performance evaluation; natural language processing

1. Introduction

In recent years, significant advancements in text-to-image (T2I) models have received increasing attention in academia and across the broader society. Many companies apply T2I models to diverse tasks, including packaging design and product launch promotions. Text–image alignment, i.e., the ability of a model to generate images accurately reflecting input text prompts, is a key capability that has attracted significant research attention [1–3]. Research on evaluating text–image alignment in T2I models is essential for assessing model performance and driving progress in the field. It is also critical for facilitating related technologies, such as text-to-video and text-to-3D generation.

The text–image alignment metric (TIAM) [4] is a template-based evaluation method specifically designed to assess text–image alignment, focusing on issues such as catastrophic neglect and attribute binding. As demonstrated by TIAM, precise control of the template structure allows text–image alignment to be evaluated in a fine-grained manner, enabling



Academic Editor: Xinming Zhang

Received: 17 December 2025

Revised: 22 January 2026

Accepted: 27 January 2026

Published: 2 February 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

separate assessment of the generation of objects and their attributes, such as color. Furthermore, by utilizing an object recognition model for accurate localization, TIAM achieves a combined fine-grained and high-precision evaluation. However, the evaluation of TIAM is not comprehensive due to two factors. First, similar to some previous studies [5,6], TIAM relies on pretrained object recognition models restricted to a narrow set of object categories. Second, although TIAM proposes an effective method for color discrimination, it cannot evaluate other attributes. Therefore, T2I models that exhibit good generation performance with such specific classes or colors are more likely to obtain higher scores. This approach risks compromising the fairness of the evaluation metric, especially for Stable Diffusion (SD) models [7], which generate images from diverse text prompts. In other words, the evaluation of SD models using TIAM is partial and incomplete.

In this study, we extend TIAM to evaluate each object in generated images, along with its attributes, actions, and positions, to provide a more comprehensive evaluation of the performance of SD models. We refer to this extended method as the Advanced Text–Image Alignment Metric (TIAM-ADV).

We focus on these aspects because they represent critical visual factors. They indicate how well a model captures the presence of objects and their properties, behaviors, and spatial relationships, all of which are essential for accurate text–image alignment. To this end, a practical method is to leverage vision–language models (VLMs) for evaluating T2I models. VLMs, such as BLIP-2, enable broad and detailed assessment of text–image alignment and are commonly employed in the evaluation of T2I models [8–11]. Nevertheless, accurately assessing all fine-grained details of the generated images remains challenging. These difficulties arise from the inherent complexity of visual scenes and the limitations in model capacity or resolution. Consequently, determining which visual details have been correctly captured becomes difficult, thereby introducing errors when assessing text–image alignment. To address these failures, we develop a method that can extract partial image regions associated with text tokens to improve the accuracy of VLMs and ensure comprehensive coverage. Our approach leverages attention maps to extract specific image regions associated with individual text tokens and feed these regions into a VLM. Attention maps effectively act as a bridge between tokens and their corresponding image regions, as shown in recent studies [12,13]. For example, prompt-to-prompt image editing [14] allows users to edit images according to text prompts by manipulating attention maps, demonstrating the ability of attention maps to localize image regions corresponding to specific tokens. Building on this capability, our previous study [15] validated the reliability of this approach for static entities (objects and attributes). The present work generalizes the framework to relational or abstract concepts (actions and positions) by leveraging their visual anchoring to associated objects, enabling a more comprehensive and fine-grained assessment of generated scenes.

Figure 1 illustrates the evaluation process of TIAM-ADV. First, a set of prompts is generated from predefined templates. Next, the target SD model generates images according to these prompts. Finally, the generated images are evaluated in terms of text–image alignment, and scores are calculated accordingly.

Regarding the experiments, we first evaluated the effectiveness of TIAM-ADV. The results show that TIAM-ADV aligns more closely with human assessments than existing methods across all aspects, demonstrating its high precision and comprehensive capability as an approach for evaluating text–image alignment. Subsequently, by applying TIAM-ADV to SD1.4, SD2, and SD3.5, we performed a detailed analysis of the characteristics of each model. The findings reveal that SD3.5 achieves greater expressiveness than SD1.4 and SD2 even though challenges in text–image alignment persist in complex tasks such as multi-attribute and multi-action generation.

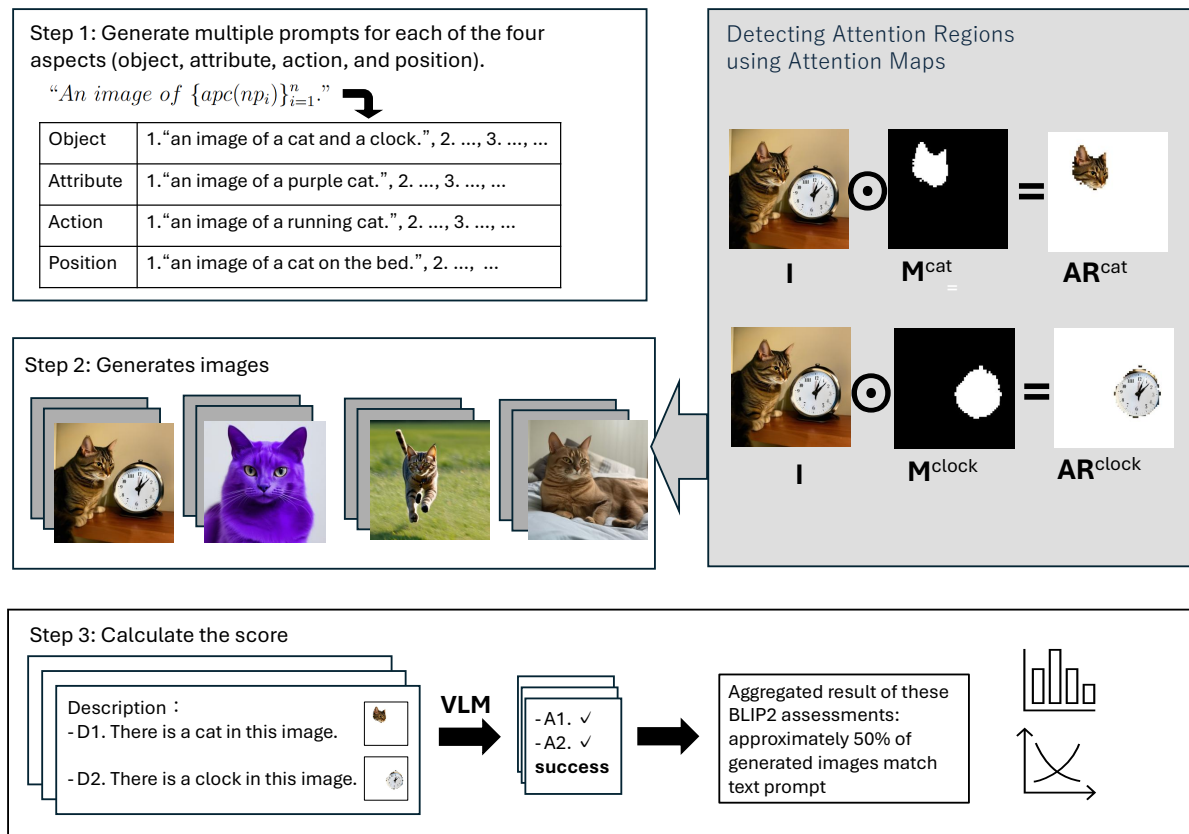


Figure 1. Overview of evaluation pipeline. (1) A set of prompts is generated from predefined templates. (2) The target SD model generates images according to these prompts. The gray image on the right illustrates how the attention region is detected using attention maps. Here, I denotes the input image, M^{cat} and M^{clock} denote the mask for the ‘cat’ and ‘clock’ tokens, respectively, and AR^{cat} and AR^{clock} denote the corresponding attention regions. (3) Calculate the TIAM-ADV score.

The primary contributions of this study are summarized as follows.

- We extend a template-based evaluation framework to cover the generation of objects, attributes, actions, and positions.
- We incorporate attention maps and VLMs into the evaluation process, enabling a more precise and comprehensive evaluation across objects, attributes, actions, and positions.
- We use TIAM-ADV to analyze the generation ability of three SD models (i.e., SD1.4, SD2, and SD3.5). The results demonstrate that SD3.5 exhibits significantly higher performance than the other two models. However, limitations in text–image alignment remain evident for more complex tasks, such as multi-attribute or multi-action generation, suggesting that further research is required to address these challenges.

2. Related Work

2.1. Evaluation of Synthetic Images

The evaluation of synthetic images has long been a critical research challenge. Early work [16] categorized evaluation metrics into two categories: plain image-based and text-conditioned metrics. Plain image-based metrics, such as the inception score (IS) [17] and Fréchet inception distance (FID) [18], primarily assess the similarity between generated and real images, without adopting any text condition. These metrics were sufficient for earlier generation models, such as generative adversarial networks [18], which exhibit a limited ability to express complex text prompts. With the rapid development of T2I models

capable of handling arbitrary text prompts, researchers have highlighted that conventional plain image-based metrics are inadequate for evaluating such models [19]. In contrast, text-conditioned metrics, which emphasize the consistency between the generated image and the input text prompt, have attracted considerable interest. Various text-conditioned evaluation methods have been proposed to assess text–image alignment [4,10,11,20].

Motivated by the need for robust evaluation, TIAM introduced a template-based approach for assessing text–image alignment in T2I models. In this work, we extend TIAM to evaluate the generation of objects, attributes, actions, and positions. This extension enables us to more comprehensively assess modern T2I models, such as stable diffusion models, across diverse application contexts.

2.2. Attention Maps in Diffusion Models

In SD models, attention maps are acquired during the image generation process. Here, we use the architectures of SD1.4 and SD2 as examples for explanation. First, a latent variable $\mathbf{Z}_T \sim N(0, I)$ is used as an initial value, and at each time step $t (t = 1, 2, \dots, T)$, noise is removed in a step-by-step manner to obtain a clean image. Here, noise removal is performed using the U-Net architecture [21], whereas the text prompt $\mathbf{p}$ is transformed into embeddings using a text encoder. The text embeddings are then fused with the visual feature embeddings at each layer of the U-Net via a cross-attention mechanism. This mechanism comprises two primary components: the query ($\mathbf{Q}$) and key ($\mathbf{K}$) matrices. The query represents the information being sought, and the key represents the reference against which the query is compared. Cross-attention uses the query and key matrices to focus on relevant information between different sources. The data from the image are projected onto the query matrix, and the data from the text prompt are projected onto the key matrix. Let $\phi(\mathbf{Z}_t)$ denote the intermediate representation of the latent variable $\mathbf{Z}_t$ at time step t within U-Net.

The query matrix is defined by

$$\mathbf{Q} = l_q(\phi(\mathbf{Z}_t)). \quad (1)$$

Here, l_q denotes a linear transformation. The resulting query matrix $\mathbf{Q} \in \mathbb{R}^{d_s \times d_e}$, where d_s denotes the dimensions of the deep spatial features and d_e denotes the dimensions of the image embedding vector space.

The key matrix is defined by

$$\mathbf{K} = l_k(\psi(\mathbf{p})). \quad (2)$$

Similar to the query matrix, l_k denotes a linear transformation. The key matrix $\mathbf{K} \in \mathbb{R}^{S \times d_e}$, where S denotes the sequence lengths (i.e., the length of the prompt after tokenization), and d_e represents the dimension of the text embedding vector space.

Then, the spatial attention maps for the prompt $\mathbf{p} = \{t_r | r \in [1, R]\}$ are obtained as follows:

$$\mathbf{AM}_{raw}^p = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_e}}\right), \quad (3)$$

where $1/\sqrt{d_e}$ denotes the scaling factor.

In U-Net, it is possible to obtain $\mathbf{AM}_{raw}^p$ at different resolutions from the layers. Following the approach outlined in the literature [14], this study considers the averaged values of $\mathbf{AM}_{raw}^p$ at a 16×16 resolution as the spatial attention maps $\mathbf{AM}^p$. Then, the spatial attention maps are resized to the image size. We denote the spatial attention maps for token t_r as $\mathbf{AM}^{t_r}$. These spatial attention maps represent the attention regions within the image corresponding to each token. Several previous studies [14,22–24] have used this

characteristic to propose image editing techniques or to enhance text–image alignment by manipulating the attention maps.

We propose a method that leverages attention maps to identify partial images of tokens and inputs them into a VLM to determine whether these tokens are accurately represented in these images. In addition, our approach can be applied to any T2I model that outputs attention maps. For example, by leveraging Python (3.11.11) libraries such as Attention Map Diffusers [25], the proposed method can be extended to other models (e.g., Flux-Schnell [26]).

2.3. Vision–Language Model

Vision–language models (VLMs) are trained on large-scale image–text pairs to extract features from images and text, and to align them in a shared or comparable embedding space. This enables cross-modal assessment of semantic similarity between images and text, and makes VLMs widely used for tasks such as visual question answering [27] and image–text matching [28]. Bootstrapping language-image Pre-training (BLIP) [8] is a foundation VLM that supports image captioning, image–text retrieval, and visual question answering tasks. BLIP-2 [9] further enables instruction-following capabilities through a two-stage pretraining strategy that introduces a learnable Q-Former to bridge the gap between a frozen image encoder and a large language model (LLM), achieving state-of-the-art performance across multiple multimodal tasks.

Building on BLIP/BLIP-2, several evaluation metrics have been proposed to quantitatively assess text–image alignment, including BLIP-2-ITC and BLIP-2-ITM [9], DA-score [10], and T2I-CompBench++ [11]. These methods pass the entire image to BLIP-2 for evaluation or scoring. In contrast, to improve accuracy, our approach passes part of the image to BLIP-2 for evaluation.

3. Materials and Methods

In this section, we introduce TIAM-ADV, a method for evaluating the performance of SD models across four aspects of their generated images: object, attribute, action, and position. Figure 1 shows an overview of TIAM-ADV. For a given model under evaluation, we first generate multiple prompts corresponding to the four aspects. Based on these prompts, the target model generates images for performance assessment. For each image, our method computes an evaluation score using attention maps and a VLM, which are then aggregated into an overall performance score.

3.1. Structuring Prompt Templates

The template-based text generation method [29] constructs predefined templates and generates sentences by filling in the slots with required information. TIAM leverages this approach, demonstrating that its key advantage lies in providing quantitative evaluations across diverse aspects. Inspired by TIAM, we define the following basic prompt template:

$$“An image of \{apc(np_i)\}_{i=1}^n.” \quad (4)$$

The template contains n noun phrases np_i . $apc(np_i)$ denotes a function that systematically attaches linguistic modifiers, such as articles, prepositions, and conjunctions, to the noun phrase np_i to improve the fluency of the resulting sentence.

A prompt is composed of multiple noun phrases np_i , which are defined as follows.

$$np_i = \left\{ \underbrace{attr_i | act_i}_{\text{Attribute or Action}} \quad \underbrace{obj_i}_{\text{Object}} \quad \underbrace{pos_i \quad obj_i^{pos_i}}_{\text{Position}} \right\}. \quad (5)$$

Each noun phrase consists of four parts: (1) an object $obj_i \in OBJ_i$, (2) an attribute $attr_i \in ATTR_i$ modifying the object, (3) an action $act_i \in ACT_i$ performed by the object, and (4) a position $pos_i \in POS_i$ indicating the spatial position of the object with another object $obj^{pos_i} \in OBJ^{pos_i}$. Among the attribute, action, and position aspects, at most one may be selected, and it is also permissible to select none. The reason is that the generation ability under the three aspects is typically evaluated independently to obtain clear and explainable experimental results. Rather than complicate prompts that encompass them all, we use only one aspect to generate prompts, which is predefined before evaluation experiments.

For example, to generate text prompts from the template, let $n = 2$ and assign each part of the noun phrase to the following sets of words: $OBJ_1 = OBJ_2 = \{ 'cat', 'banana' \}$, $POS_1 = POS_2 = \{ 'near', 'under' \}$, and $OBJ^{pos_1} = OBJ^{pos_2} = \{ 'desk' \}$. Our approach can produce prompts such as “an image of a cat under the desk and a banana near the desk.”

We generate all possible combinations based on the set of words used; therefore, the number of generated prompts $|P|$ can be calculated as follows:

$$|P| = \prod_{i=1}^n (|X_i| \times |OBJ_i|), \quad X_i = \{ATTR_i, ACT_i, POS_i\}. \quad (6)$$

For P , we first filter out those prompts that contain repeated words and then employ a LLM to remove prompts with unnatural language expressions.

3.2. Detecting Attention Regions Using Attention Maps

The gray area in Figure 1 illustrates an example of the attention region detection process using attention maps. Here, the white region in the mask (denoted as $\mathbf{M}^{cat}$ and $\mathbf{M}^{clock}$), obtained from the attention maps, highlights the area most strongly influenced by the corresponding token during image generation. In other words, this region of the image best represents the corresponding token. We extract and denote it as the attention region $\mathbf{AR}^{tr}$ and pair it with the corresponding token t_r .

The attention region $\mathbf{AR}^{tr}$ is computed as follows. First, we generate the attention map $\mathbf{AM}^{tr}$ corresponding to the token t_r . Specifically, we utilize the cross-attention mechanism, extracting the attention weights that align intermediate visual features with the text embedding of t_r . To ensure robustness, we collect these attention weights from selected modules of the model (e.g., those with an intermediate spatial resolution such as 16×16 in SD 1.4) and average them across these modules to form a 2D attention map. This aggregated map $\mathbf{AM}^{tr}$ represents the spatial influence of the token on the generated image. The detailed formulation and implementation are provided in Section 2.2.

The attention map is normalized to the range $[0, 1]$ and thresholded as follows:

$$\mathbf{AM}_{w,h}^{tr} = \begin{cases} 1 & \text{if } \mathbf{AM}_{w,h}^{tr} \geq 1/\theta \\ 0 & \text{else.} \end{cases} \quad (7)$$

Here, $w \in [1, W]$ and $h \in [1, H]$ denote the spatial indices along the image width and height, respectively. The threshold θ directly affects the resulting attention region extracted from the attention map. Larger values allow more pixels to be included, whereas smaller values produce more compact areas. The detailed analysis regarding the choice of θ is provided in Section 4.1.3.

Next, we focus on the token $t_r \in \{attr_i, act_i, obj_i, pos_i\}$ within a noun phrase (np_i). We then elaborate on the methods for obtaining the mask $\mathbf{M}^{tr}$ corresponding to each token t_r .

For a token $t_r \in \{obj_i, attr_i, act_i\}$,

$$\mathbf{M}^{tr} = \mathbf{AM}^{obj_i} \vee \mathbf{AM}^{tr}. \quad (8)$$

For a token $t_r = pos_i$, because the spatial position refers to the relationship between two objects, the corresponding mask is obtained by aggregating the maps of both objects.

$$\mathbf{M}^{t_r} = \mathbf{AM}^{obj_i} \vee \mathbf{AM}^{pos_i} \vee \mathbf{AM}^{obj^{pos_i}} \quad (9)$$

where $\vee$ denotes element-wise logical OR for binary attention masks. We use the logical OR to safely include all relevant regions, even if it slightly enlarges the attention area.

The design of mask generation is motivated by the observation that, in image generation, attributes, actions, and positions are typically manifested through and anchored to the associated objects. As a result, the attention maps of such tokens tend to concentrate on the relevant objects and their surrounding regions. Therefore, we apply thresholding to isolate strong attention signals and integrate the associated object attention map to capture more comprehensive information. We empirically validate its reliability for practical applications in Section 4.2.

Finally, with the original image denoted by $\mathbf{I}$, the attention region $\mathbf{AR}^{t_r}$ is computed as follows:

$$\mathbf{AR}^{t_r} = \mathbf{I} \odot \mathbf{M}^{t_r}, \quad (10)$$

where $\odot$ denotes element-wise multiplication. We found that setting the background of $\mathbf{AR}^{t_r}$ to white enhances the correlation with human judgments. Therefore, the black background is converted to white by inverting the mask $\mathbf{M}^{t_r}$ and applying it via pixel-wise addition.

3.3. Designing Scores

Following previous research [4], for the target model under evaluation, our method computes a score as the average over all generated images. BLIP-2 was employed to determine the evaluation value for each image. The formula is defined as follows:

$$E_{\mathbf{p} \in P, \chi \sim N(0, I)} \left[\prod_{i=1}^n \text{Step}(\text{BLIP2}(\text{SD}(\mathbf{p}, \chi) \odot \mathbf{M}^{t_r}, D_{t_r})) \right],$$

$$t_r \in \{obj_i, attr_i, act_i, pos_i\} \quad (11)$$

Let P denote the set of prompts generated by the templates. For each $\mathbf{p} \in P$, the images are generated using different $\chi \sim N(0, I)$ sampled with different seeds. Here, n denote the number of np_i , and $\text{SD}(\mathbf{p}, \chi) \odot \mathbf{M}^{t_r}$ (i.e., $\mathbf{AR}^{t_r}$) represents the partial image corresponding to the token t_r . In addition, the function D_{t_r} generates the descriptive text to be input into BLIP-2 and is defined as follows:

$$\begin{aligned} D_{obj_i} &= \text{"There is } apc([obj_i]) \text{ in this image."} \\ D_{attr_i} &= \text{"There is } apc([attr_i][obj_i]) \text{ in this image."} \\ D_{act_i} &= \text{"There is } apc([act_i][obj_i]) \text{ in this image."} \\ D_{pos_i} &= \text{"There is } apc([obj_i][pos_i][obj^{pos_i}]) \text{ in this image."} \end{aligned} \quad (12)$$

Similar to Equation (4), $apc(\cdot)$ denotes the function that adds linguistic modifiers to improve sentence fluency. For example, for the given prompt "an image of a big cat" the function D_{attr_i} generates the following descriptive text: "There is a big cat in this image." Along with the corresponding attention region, the generated descriptive text forms what we refer to as a region–text pair, which serves as the basic evaluation unit in our framework. BLIP-2 then returns a probability value that measures the consistency of the region–text pair, indicating how well the descriptive text aligns with the corresponding image region. We apply a step function that compares the probability with a threshold τ to obtain a

binary decision: if the probability is greater than or equal to the threshold τ , the pair is considered consistent (success, returning 1); otherwise, it is considered inconsistent (failure, returning 0). Section 4.1.3 describes the procedure for determining the threshold values. An image generation process is considered successful only when the image and text prompt fully match, i.e., when all corresponding descriptive text receives a value of 1. After averaging the values over multiple images to obtain a score, a higher score indicates that the images better match the prompts and implies improved model performance.

4. Experiments and Result Analysis

This section first describes the implementation details in Section 4.1. To assess the effectiveness of TIAM-ADV, Section 4.2 presents a user study conducted to compare its results with human evaluations. Moreover, we analyze the application of TIAM-ADV to three models, i.e., SD1.4, SD2, and SD3.5. Their characteristics are summarized in Section 4.3.

4.1. Implementation Details

4.1.1. Dataset

Selecting an appropriate predefined dataset is crucial for a comprehensive and effective evaluation. To address a more general case, we constructed a dataset using the process described below and summarized in Figure 2.

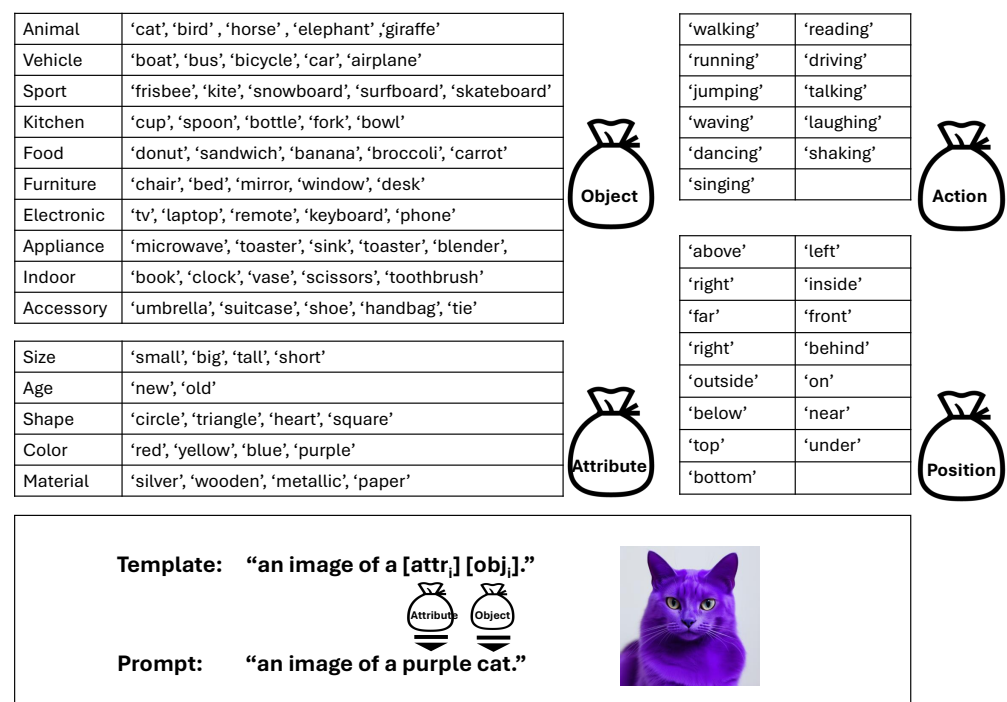


Figure 2. Summary of predefined dataset.

- **Object set:** We randomly selected 10 super categories from the common objects in context dataset [30] and sampled five object categories from each super category, resulting in an object set of 50 object categories (words).
- **Attribute set:** We referred to Practical English Usage [31] and A Practical English Grammar [32] to identify the major categories of adjectives. Consequently, the major categories include size, age, shape, color, and material, and the adjective set comprises 18 commonly used adjectives.

- Action set: We selected 11 dynamic verbs from the Oxford 3000/5000 list [33], including ‘walking,’ ‘running,’ ‘jumping,’ ‘waving,’ ‘dancing,’ ‘singing,’ ‘reading,’ ‘driving,’ ‘talking,’ ‘laughing,’ and ‘shaking.’ These verbs were selected for three reasons: (1) high frequency in daily use, (2) visual distinctiveness, enabling reliable recognition in generated images, and (3) natural co-occurrence with objects of interest, specifically, animals from our object set and the single human category “person,” to minimize the likelihood of unnatural prompts in the action-related experiments.
- Position set: Following existing research [34], we selected adverbs describing the relative positions between the two objects to construct the following position set: {‘above’, ‘right’, ‘far’, ‘outside’, ‘below’, ‘top’, ‘bottom’, ‘left’, ‘inside’, ‘front’, ‘behind’, ‘on’, ‘near’, ‘under’}.

4.1.2. Sampling Based Evaluation Protocol

For the generated prompts P , we employed a LLM to filter out prompts with unnatural language expressions and then classified the remaining prompts into two categories: realistic and nonrealistic. Although the filtering and classification process eliminated a significant portion of the prompts, the theoretical cardinality of the full set P (Equation (6)) remained extremely large, making exhaustive evaluations impractical. To make the evaluation feasible, we adopted random sampling to estimate model performance scores.

To balance the trade-off between accuracy and computation, we determine the required size based on an acceptable error margin and a 95% confidence interval. For example, to achieve a ± 0.05 error margin under a conservative assumption, approximately 385 randomly sampled prompts per template are sufficient. This approach ensures reliable evaluation scores without exhaustively evaluating the entire set.

In this experiment, we begin with approximately 300 randomly sampled prompts from the entire set. Following prior work [4], this sample size is sufficient to obtain reasonably accurate estimates of model performance. We additionally compute the corresponding error margins to quantify uncertainty. To enable performance comparisons across models and support statistical significance testing, we further increase the number of sampled prompts as required by the specific analysis.

4.1.3. Threshold Determination

We determine thresholds for two distinct purposes in our study: one for binarizing the attention maps (θ) and another for binarizing the BLIP2 predictions (τ).

Attention Map Threshold θ

To determine an appropriate threshold θ for binarizing the attention map, we first conducted a qualitative inspection by visualizing attention regions generated with different threshold values and comparing their effects intuitively. The threshold was explored in the range of 1 to 10 with a step size of 0.5. Several representative visualization examples are shown in Figure 3. As θ increases, the attention regions evolve from failing to adequately cover critical areas, to accurately retaining the relevant areas, and eventually to including an increasing amount of irrelevant areas. Empirically, we observe that $\theta = 3.5$ provides a balanced trade-off between retaining critical areas and removing irrelevant areas.

In addition to qualitative inspection, we further performed a quantitative analysis to support the selection of θ . Specifically, we randomly sampled images generated by SD1.4, SD2, and SD3.5 from the experiment described in Section 4.3, and manually annotated the image areas relevant to the corresponding tokens. This process resulted in a total of 60 human-annotated ground-truth masks. For each image, predicted masks were obtained from the attention map under different θ values, and the Intersection over Union (IoU) was computed between each predicted mask and its corresponding human annotation. The

mean IoU (mIoU) [35] was then calculated by averaging IoU values over all annotated images. Considering the substantial cost of manual annotation and the fact that this analysis aims to examine relative mIoU trends rather than absolute performance, we limited the evaluation to 60 annotated samples, which we found sufficient to observe consistent trends. As shown in Figure 4, the mIoU reaches its peak around $\theta = 3.5$, where the predicted masks exhibit the closest alignment with the human annotations. Based on both qualitative and quantitative evidence, we select $\theta = 3.5$ as the attention map threshold in all subsequent experiments.

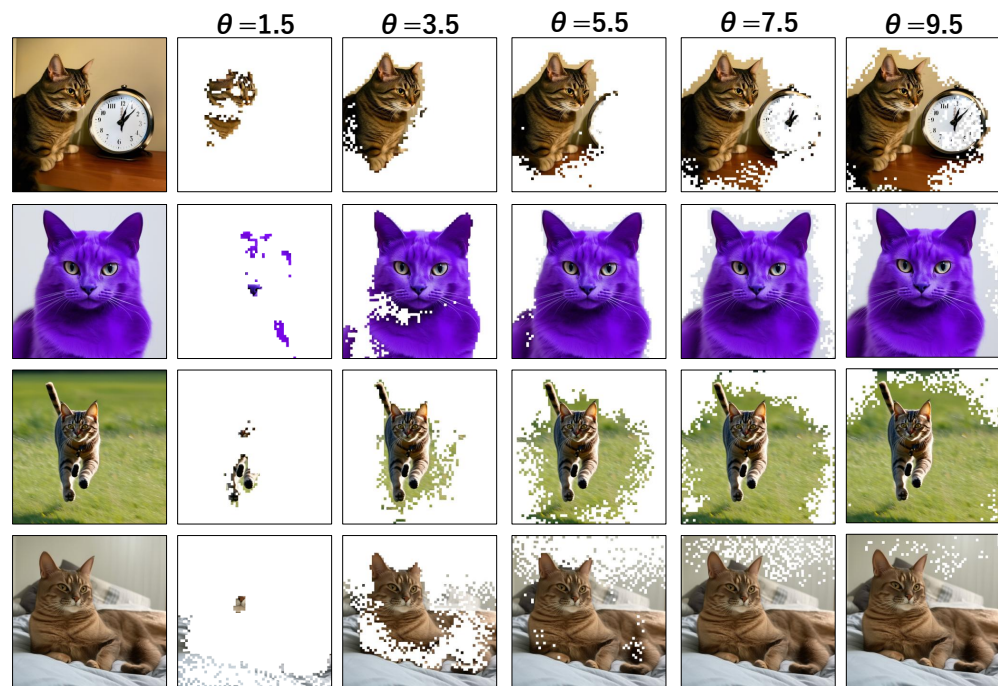


Figure 3. Visualization of attention regions under different threshold values θ .

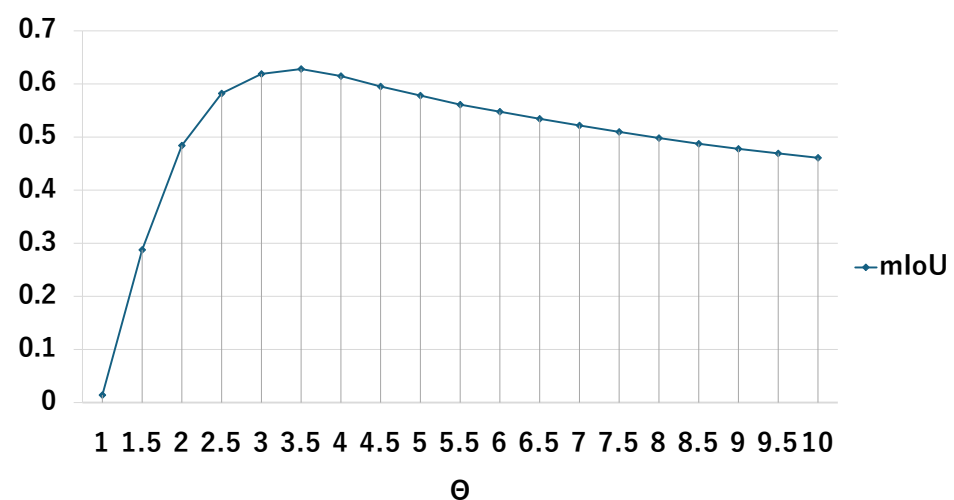


Figure 4. mIoU under different threshold values θ .

BLIP2 Prediction Threshold τ

To select an appropriate threshold τ for binarizing BLIP2 predictions, we evaluated all 180 region–text pairs, which were balanced across the three models and all templates. The procedure was as follows:

1. Receiver operating characteristic (ROC) curves were constructed for each evaluation aspect.
2. We calculate Youden's index for the points on the ROC curves. It is defined as $J = TPR - FPR$, where TPR and FPR denote the true and false positive rates, respectively.
3. The optimal threshold τ was explored in the range of 0.25–0.85 with a step size of 0.05. The selected optimal threshold was then used to maximize the J value.

Consequently, the optimal thresholds τ were determined as 0.45, 0.55, 0.55, and 0.65 for objects, attributes, actions, and positions, respectively.

4.1.4. Implementation

All experiments in this study were conducted using PyTorch (2.7.1). Token-level attention maps were obtained following previous studies, such as prompt-to-prompt [14] and attention-map-diffusers [25]. The experiments were run on SQUID (Supercomputer for Quest to Unsolved Interdisciplinary Data Science). We selected SD1.4, which employs a U-Net + VAE + CLIP architecture, and SD2, which replaces CLIP with OpenCLIP to improve text–image alignment. SD3.5 further advances text–image alignment by incorporating a large language encoder (T5-XXL) and more flexible CLIP variants, together with the MMDiT Transformer generator and training techniques such as QK normalization, enabling improved understanding and execution of complex prompts. The number of sampling steps was fixed to 50 for SD1.4 and SD2 and 15 for SD3.5, as sampling steps are not directly comparable due to the rectified flow formulation in SD3.5, which enables near-converged image quality with fewer steps, ensuring a comparable convergence regime for each model. The classifier-free guidance scale was set to 7.5. Prompt checking and classification were performed using Gemini 2.5 Pro [36].

4.2. User Study

In this experiment, we aimed to measure the gap between human judgment and the scores produced by various evaluation metrics, including the proposed method and existing approaches. First, we conducted a T2I matching survey of the generated images with 30 human participants. The survey presented the participants with several prepared images, each accompanied by a descriptive text generated by D_{tr} . The participants were asked to place a check mark before the text if the image matched it. As a restriction, the participants were instructed to ignore aesthetics or personal preference, and to focus solely on the text–image alignment of each image. Figure 5 shows some example questions from the survey in Google Forms. We used 120 images generated in the experiment described later in Section 4.3. Table 1 summarizes the object, attribute, action, and position aspects of these images, including the number of erroneous images identified in each aspect. The results of the participants yielded Fleiss' kappa of 0.694, indicating “substantial consistency” [37] confirming the reliability of the survey results.

Table 1. Summary of properties of 120 evaluated images. Numbers in parentheses indicate the number of erroneous images in each aspect.

Aspect	Number of Images (Erroneous)
Object	30 (10)
Attribute	30 (14)
Action	30 (17)
Position	30 (14)
Total	120 (55)

Queries about object generation
☐ There is a boat in this image.

☐ There is a chair in this image.
Queries about attribute generation
☐ There is a big cat in this image.

☐ There is a wooden chair in this image.
Queries about action generation
☐ There is a running horse in this image.

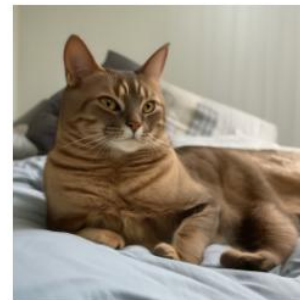
☐ There is a singing person in this image.
Queries about Position generation
☐ There is a cat on the bed in this image.

Figure 5. User interface. We asked the participants to place a check mark before the descriptions that match the image. There may be multiple correct answers. The descriptions for each image were derived from its prompt using Equation (12). The prompts were: “An image of a boat and a chair.” (**top-left**), “An image of a big cat and a wooden chair.” (**top-right**), “An image of a running horse and a singing person.” (**bottom-left**), “An image of a cat on the bed.” (**bottom-right**).

Second, we compared the human judgment with the scores produced by BLIP-2, PickScore [38], ImageReward [39], HPS v2 [40], and TIAM-ADV using Pearson correlation to quantify how closely each automatic metric reflects human judgment. BLIP-2, PickScore, ImageReward, and HPS v2 are commonly used metrics for evaluating image–text alignment and visual quality in T2I models. Table 2 presents the experimental results. Compared with previous studies, the Pearson correlation coefficient of TIAM-ADV was the highest, reaching 0.853 ($p < 10^{-48}$). Typically, a correlation greater than 0.7 is considered a strong positive association, indicating that the proposed method aligns closest with human judgment.

Table 2. Pearson correlation coefficients for each method. The highest value is highlighted in **bold**.

BLIP2	Pickscore	ImageReward	HPS v2	TIAM-ADV
0.504	0.620	0.823	0.712	0.853

The prediction capabilities of the metrics were further assessed using ROC curves, as shown in Figure 6. Among them, TIAM-ADV achieved the highest predictive performance, with an area under curve (AUC) of 0.96, slightly exceeding ImageReward and outperforming other metrics. Unlike ImageReward, our method does not require additional training of the VLM, providing greater flexibility. Notably, TIAM-ADV also outperformed BLIP-2 (AUC 0.82) and exhibited a stronger correlation with human judgment, highlighting the effectiveness of its attention region extraction of tokens, which distinguishes it from BLIP-2. This confirms that the attention maps are sufficiently informative for the evaluation task considered in this work.

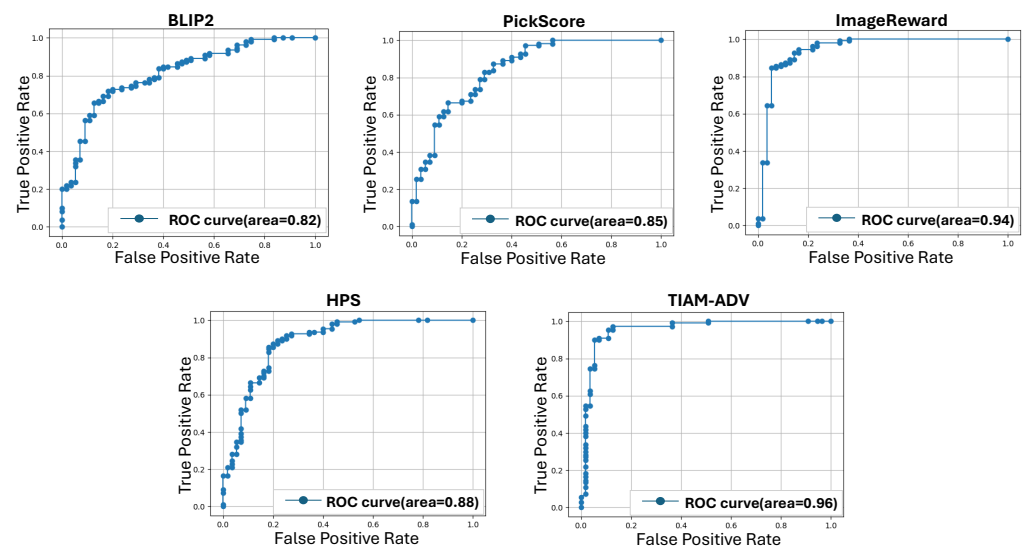


Figure 6. ROC curves, with the corresponding AUC values annotated within each figure.

Overall, TIAM-ADV exhibits the highest correlation with human judgment, demonstrating its effectiveness for assessing text–image alignment. A brief comparison of the other metrics clarifies their complementary roles. As highlighted in previous research, PickScore predicts relative scores with an emphasis on user preferences, HPS v2 provides absolute scores but primarily emphasizes image quality rather than text–image alignment, and ImageReward considers both alignment and factors such as fidelity and harmlessness. In contrast, TIAM-ADV generates absolute scores solely based on text–image alignment, making it particularly reliable when alignment is the primary concern. At the same time, it can still serve as a complement to such approaches. Meanwhile, TEM-ADV does not require any additional training and can readily utilize different VLMs as needed, making it a flexible and lightweight evaluation approach.

4.3. Results of Model Comparison

In this experiment, we adapted our method to three models, namely, SD1.4, SD2, and SD3.5. Our method evaluated their performance by assigning scores using four prompt sets that focused on the image generation of objects, attributes, actions, and positions. These prompt sets were generated from templates (Table 3) and word sets (Figure 2). Based on these scores, we further analyzed the differences in performance among the three models and identified their individual characteristics.

Table 3. Overview of the templates used for assessing object, attribute, action, and position generation.

Aspect	Templates Used in Experiments
Object	Single: “an image of obj_i .” Multiple: “an image of obj_i and obj_j .” ($obj_i \neq obj_j$)
Attribute	Single: “an image of $attr_i$ obj_i .” Multiple: “an image of $attr_i$ obj_i and $attr_j$ obj_j .” ($obj_i \neq obj_j$ and $attr_i \neq attr_j$)
Action	Single: “an image of act_i obj_i .” Multiple: “an image of act_i obj_i and act_j obj_j .” ($obj_i \neq obj_j$ and $act_i \neq act_j$)
Position	Single: “an image of $apc([obj_i] [pos_i] [obj^{pos_i}])$.”

First, we report the performance scores for object generation. The *Object* set consists of 50 object categories. For each distinct pair of categories (obj_i, obj_j) with $obj_i \neq obj_j$, we constructed the prompt “an image of obj_i and obj_j ” (controlling the articles as expected), generating 1225 images in total. We randomly sampled 600 images each from the images generated by SD1.4 and SD2 and 300 images from those generated by SD3.5 to ensure sufficient data for statistical significance testing. The alignment between the prompts and the images was then evaluated using TIAM-ADV. For comparison, we also experimented with 300 images, which were generated from the simpler prompt “an image of obj_i ,” while changing the random seed χ .

The results are presented in Table 4, which reports the evaluation scores for object generation using prompts with one or two objects. The values in parentheses indicate the error margin and the 95% confidence interval. All models performed well in generating images containing a single object, achieving scores above 0.847. However, when generating two objects simultaneously, SD1.4 and SD2 struggled, achieving significantly lower scores of 0.338 and 0.399, respectively. In contrast, SD3.5 consistently maintained a score of 0.904. Significance analysis confirms that the differences between the three models are statistically significant at the 95% confidence level. These findings indicate that although SD1.4 and SD2 exhibit comparable performance, SD2 performs slightly better in generating images with two objects. Overall, SD3.5 substantially outperforms both models.

Table 4. Evaluation score for object generation.

Model	(obj_1)	(obj_1, obj_2)
SD1.4	0.920	0.338 (95% CI 0.311–0.365)
SD2	0.847	0.399 (95% CI 0.371–0.427)
SD3.5	0.954	0.904 (95% CI 0.875–0.933)

Next, we characterized the capacity of the models to apply attributes to objects. Two types of prompts were considered: “an image of $attr_i$ obj_i ” and “an image of $attr_i$ obj_i and $attr_j$ obj_j ,” where $obj_i \neq obj_j$ and $attr_i \neq attr_j$, which were obtained from the *Object* and *Attribute* sets. For the single-attribute prompts, 900 (18×50) prompts were generated. Using a LLM, we filtered out prompts exhibiting unnatural language expressions and classified the remaining prompts into “realistic” and “unrealistic” categories. We then randomly sampled 300 images from each category. For the two-attribute prompts, only the realistic prompts were considered. To ensure a sufficient number of samples for statistical significance testing, we randomly sampled 600 images each from the images generated by SD1.4 and SD2 and 300 images from those generated by SD3.5.

The results are presented in Table 5, where the columns correspond to prompts with one realistic attribute, one unrealistic attribute, and two realistic attributes. For the single-attribute prompts, SD1.4 and SD2 achieved scores of approximately 0.7, whereas

SD3.5 achieved a score of 0.877. The scores of SD1.4 and SD2 dropped to 0.035 and 0.068, respectively, when generating two objects with attributes. In contrast, SD3.5 consistently maintained a score of 0.625 for the two-attribute prompts. The statistically significant difference indicates that SD1.4 performs worse than SD2, which is outperformed by SD3.5 in multi-attribute generation. All models did not exhibit a significant difference between the realistic and unrealistic categories. Compared with object-only generation (Table 4), adding attribute generation led to decreased scores across all models. These results suggest that although SD1.4 and SD2 exhibit comparable performance for single-attribute prompts, SD2 outperforms SD1.4 in multi-attribute generation, and SD3.5 substantially outperforms both in terms of attribute generation.

Table 5. Evaluation score for attribute generation.

Model	$attr_1$ (Realistic)	$attr_1$ (Unrealistic)	$(attr_1, attr_2)$ (Realistic)
SD1.4	0.700 (95% CI 0.663–0.737)	0.699	0.035 (95% CI 0.020–0.05)
SD2	0.680 (95% CI 0.643–0.717)	0.637	0.068 (95% CI 0.048–0.088)
SD3.5	0.877 (95% CI 0.851–0.903)	0.885	0.625 (95% CI 0.570–0.680)

To evaluate the ability of the models to associate actions with objects, we used the set *Action* to construct prompts. To avoid unnatural prompts, the object set was restricted to $OBJ_i = \{\text{'bird'}, \text{'cat'}, \text{'horse'}, \text{'elephant'}, \text{'giraffe'}, \text{'person'}\}$. Specifically, two types of prompts were considered: “an image of $act_i\ obj_i$ ” and “an image of $act_i\ obj_i$ and $act_j\ obj_j$,” where $obj_i \neq obj_j$ and $act_i \neq act_j$. For the first prompt type, we generated 66 text prompts (including 35 unrealistic prompts such as “An image of a talking giraffe,” and 31 realistic prompts), yielding 660 images by changing the random seed χ . For the second prompt type, we randomly sampled 600 images each from the images generated by SD1.4 and SD2 and 300 images from those generated by SD3.5, all drawn from the realistic prompts.

The results (Table 6) indicate that for all three models, generation using realistic prompts outperformed that using unrealistic prompts. Furthermore, the score decreased as the number of specified actions increased. SD3.5 substantially outperformed the other two models. In particular, in the two-action generation experiment, SD3.5 achieved a score of 0.422, which was significantly higher than those of the other two models (0.003 and 0.025, respectively). Between these two models, SD2 outperformed SD1.4 in multi-action generation, with statistically significant results. Overall, more than half of the image generations involving two actions failed. Therefore, we consider that enhancing image generation for multiple actions would be a promising research direction for future work.

Table 6. Evaluation score for action generation.

Model	act_1 (Realistic)	act_1 (Unrealistic)	(act_1, act_2) (Realistic)
SD1.4	0.748	0.425 (95% CI 0.404–0.446)	0.003 (95% CI 0.002–0.004)
SD2	0.739	0.577 (95% CI 0.556–0.598)	0.025 (95% CI 0.022–0.028)
SD3.5	0.839	0.597 (95% CI 0.576–0.618)	0.422 (95% CI 0.381–0.463)

To evaluate the ability of the models to generate images depicting the spatial relationships between objects, we considered prompts of the form “an image of $apc([obj_i] [pos_i] [obj_j^{pos_i}])$,” where obj_i and pos_i are from the *Object* and *Position* sets, respectively, and $obj_j^{pos_i} \in \{\text{'chair'}, \text{'bed'}, \text{'mirror'}, \text{'door'}, \text{'desk'}\}$. We obtain 300 prompts from all possible combinations after filtering out unnatural expressions.

The experimental results are presented in Table 7, showing that SD3.5 also substantially outperforms SD1.4 and SD1.5 in representing positions, achieving a score of 0.851. Compared with object generation, adding positional relationships reduces the score of

image generation, particularly for SD1.4 and SD2, whose scores drop from 0.92 and 0.847 to 0.357 and 0.381, respectively.

Table 7. Evaluation score for position generation.

Model	<i>obj₁</i>	<i>pos₁</i>
SD1.4	0.920	0.357 (95% CI 0.303–0.411)
SD2	0.847	0.381 (95% CI 0.326–0.436)
SD3.5	0.954	0.851 (95% CI 0.811–0.891)

The experimental results reveal that the text–image alignment performance of SD1.4 and SD2 is roughly comparable. However, SD2 exhibits better performance in multi-object, multi-attribute, and multi-action image generation. SD3.5 consistently outperforms SD1.4 and SD2 across multiple generation tasks. Notably, substantial improvements were observed in position, multi-attribute, and multi-action generation. However, the most significant limitation of SD3.5 lies in its multi-action generation capability, which may represent a valuable direction for further exploration in future research.

5. Discussion

In this study, we introduce TIAM-ADV, a comprehensive, fine-grained, and high-precision method for evaluating text-to-image alignment. It enables researchers to examine model behavior across various visual aspects, establishing a more rigorous basis for future T2I model evaluation.

5.1. Theoretical Implications

TIAM-ADV extends the template-based TIAM framework by assessing multiple aspects of generated images, thereby advancing the systematic understanding of T2I alignment. Unlike prior approaches that either rely on object detection models to identify image regions corresponding to tokens or require additional training on large-scale human evaluation data to simulate human judgments, our method leverages attention maps to enhance evaluation precision, demonstrating the feasibility of using attention-based analysis for high-precision evaluation.

Furthermore, the flexibility of attention maps opens up opportunities for developing multifaceted evaluation metrics, such as layer-wise or time-step-based analyses of attention within U-Net architectures, offering new directions for theoretical exploration.

5.2. Practical Implications

TIAM-ADV provides a lightweight and flexible evaluation framework with the potential to accommodate different VLMs without additional training. This adaptability is particularly valuable for domain-specific applications. For instance, evaluating the performance of T2I models in the medical domain, such as for pathology image generation, can be achieved by integrating domain-specific VLMs, such as Pathology Language and Image Pretraining (PLIP), without retraining the evaluation framework. In contrast, existing methods (e.g., PickScore, HPS v2, ImageReward) that require additional training cannot generalize effectively to specialized domains.

Furthermore, TIAM-ADV focuses exclusively on text–image alignment, complementing other metrics that assess human preferences or aesthetics. Consequently, high TIAM-ADV scores alongside lower scores on other metrics can reveal alignment proficiency even when visual quality or aesthetic appeal is lower, enabling practitioners to interpret model behavior more accurately in real-world applications.

5.3. Limitations and Future Work

Despite its advantages, the proposed method has a few limitations.

First, selecting an appropriate dataset remains challenging. Ensuring sufficient coverage of relevant objects, attributes, actions, and positions while avoiding unnatural or ambiguous language requires careful design. To address this, we plan to leverage LLMs to assist in generating and filtering high-quality prompt sets, potentially automating parts of the dataset construction process to ensure linguistic naturalness and alignment with evaluation objectives. In addition, considering the complexity of this process, future work may incorporate multiple AI models in a collaborative framework to further enhance the robustness of the overall pipeline.

Second, our approach does not address the issue of attribute leakage. Attribute leakage occurs when attributes erroneously appear on unintended objects. For example, as shown in Figure 7, the attribute “yellow,” which should apply to the car, has instead leaked onto the walls. To quantify this limitation, we conducted a supplementary analysis using prompts following the template “an image of $attr_i$ obj_i ” across SD1.4, SD2, and SD3.5 (100 images each). Manual inspection revealed attribute leakage rates (the proportion of images where the attributes appeared in non-target regions) of 0.142, 0.167, and 0.080, respectively. Despite its occurrence, since our method evaluates only the objects specified in the prompt, it ignores background regions and thus cannot capture attribute leakage. This design, inherited from the original TIAM, aligns with user intent: for instance, given the prompt “an image of a yellow car,” a user typically considers the generation successful if the vehicle appears yellow, even if parts of the background share the same color. However, to enable stricter evaluation standards and more comprehensive scene analysis, future work will aim to extend the method to detect attribute leakage through attention map analysis. Specifically, this can be achieved by examining whether attention maps corresponding to an attribute token extend beyond the target object.



Figure 7. Example of attribute leakage. The image is generated from the prompt “An image of a yellow car.”

In addition, our method utilizes thresholded attention maps to generate binary masks. While experiments in Section 4.2 demonstrate strong agreement with human judgments, we acknowledge this strategy captures approximate token–image correspondence without explicitly distinguishing token types. Specifically, relational or abstract concepts such as

actions and positions may require distinct handling compared to objects. In future work, we plan to explore more flexible mask generation mechanisms, such as token-specific modeling strategies, to more accurately extract relevant regions and improve evaluation performance.

6. Conclusions

As an evaluation metric for T2I models, TIAM relies on predefined templates to measure consistency in relatively simple aspects such as object count and color. Although this approach is effective within its scope, it is inherently limited because it overlooks other crucial visual factors that determine how well a generated image aligns with its text prompt. In this study, we address this limitation by extending TIAM to a more advanced framework, termed TIAM-ADV. The proposed method integrates attention maps with VLMs, thereby enabling a comprehensive, fine-grained, and high-precision evaluation. Unlike the original TIAM, TIAM-ADV evaluates not only objects and colors but also attributes, actions, and positions, which are essential for capturing the semantic richness of text prompts. To validate the effectiveness of TIAM-ADV, we analyzed the evaluation scores of images generated by the proposed method and compared them with human judgments. The results demonstrate that TIAM-ADV outperforms existing methods by exhibiting stronger correlations with human judgments. Furthermore, we applied the proposed method to evaluate the generation abilities of three SD models (SD1.4, SD2, and SD3.5). The findings indicate that SD3.5 exhibits superior expressiveness compared with SD1.4 and SD2. However, limitations in text–image alignment remain evident in more complex scenarios, such as multi-attribute or multi-action generation.

Future work will focus on extending the method to address challenging text–image alignment issues, such as attribute leakage. We also aim to apply the metric to real-world scenarios where structured or template-like text is prevalent. One example is the medical domain, which is experiencing rapid growth in text-to-image (T2I) model applications [41]. We plan to leverage large language models to assist in generating high-quality prompt sets, thereby facilitating dataset construction and ensuring linguistic naturalness.

Author Contributions: Conceptualization, H.F. and C.L.; methodology, H.F., S.O. and H.S.; software, H.F.; validation, H.F.; formal analysis, H.F. and H.S.; investigation, H.F.; resources, C.L.; data curation, H.F.; writing—original draft preparation, H.F.; writing—review and editing, H.F., C.L., S.O., K.F. and H.S.; visualization, H.F.; supervision, C.L. and H.S.; project administration, C.L. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Our study did not require approval from an Ethics Committee or Institutional Review Board (IRB). The research consisted of an anonymous, non-interventional questionnaire survey on text–image alignment evaluation and involved no personal, sensitive, medical, or psychological data. According to Japanese national regulations, including the Ethical Guidelines for Life Science and Medical Research Involving Human Subjects, ethical review is required only for medical or life-science research that may affect human life, health, or dignity. As this study falls within the field of information science and posed no risk to participants, it lies outside the scope of these guidelines. This approach is consistent with the Declaration of Helsinki, which requires ethical approval only when research may affect participants' health or well-being.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study. Prior to participation, participants were informed of the purpose of the study, the voluntary nature of their participation, the anonymity of the questionnaire, and the fact that no personal, sensitive, medical, or psychological data would be collected. Completion and submission of the anonymous questionnaire were considered to constitute informed consent.

Data Availability Statement: The data used in this study were generated by the authors. Details of the data generation process are provided in Section 4.1.1. All datasets referenced during the generation process are publicly available and are cited in the manuscript.

Acknowledgments: We sincerely thank all the participants who took the time to complete our survey. Their valuable responses and insights greatly contributed to this research. This work was partly completed using SQUID at D3 Center, Osaka University.

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

Abbreviations

The following abbreviations are used in this manuscript:

T2I	Text-to-image
SD	Stable diffusion
TIAM	Text-image alignment metric
TIAM-ADV	Advanced text-image alignment metric
VLM	Vision-language model
BLIP	Bootstrapping language-image pre-training
LLM	Large language model
IoU	Intersection over union
ROC	Receiver operating characteristic
AUC	Area under curve

References

1. Liu, L.; Du, C.; Pang, T.; Wang, Z.; Li, C.; Xu, D. Improving Long-Text Alignment for Text-to-Image Diffusion Models. *arXiv* **2025**, arXiv:2410.11817.
2. Grimal, P.; Borgne, H.L.; Ferret, O. Text-to-Image Alignment in Denoising-Based Models through Step Selection. *arXiv* **2025**, arXiv:2504.17525.
3. Izadi, A.M.; Hosseini, S.M.H.; Tabar, S.V.; Abdollahi, A.; Saghafian, A.; Baghshah, M.S. Fine-Grained Alignment and Noise Refinement for Compositional Text-to-Image Generation. *arXiv* **2025**, arXiv:2503.06506.
4. Grimal, P.; Le Borgne, H.; Ferret, O.; Tourille, J. TIAM-A metric for evaluating alignment in Text-to-Image generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 2890–2899. [\[CrossRef\]](#)
5. Ghosh, D.; Hajishirzi, H.; Schmidt, L. GenEval: An object-focused framework for evaluating text-to-image alignment. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 52132–52152.
6. Hinz, T.; Heinrich, S.; Wermter, S. Semantic object accuracy for generative text-to-image synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1552–1565. [\[CrossRef\]](#) [\[PubMed\]](#)
7. 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. [\[CrossRef\]](#)
8. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
9. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
10. Singh, J.; Zheng, L. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA 10–16 December 2023; pp. 70799–70811.
11. Huang, K.; Duan, C.; Sun, K.; Xie, E.; Li, Z.; Liu, X. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 3563–3579. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; Ture, F. What the daam: Interpreting stable diffusion using cross attention. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 5644–5659. [\[CrossRef\]](#)

13. Zhang, G.; Wang, K.; Xu, X.; Wang, Z.; Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 16–22 June 2024; pp. 1755–1764. [\[CrossRef\]](#)
14. Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; Cohen-Or, D. Prompt-to-Prompt Image Editing with Cross-Attention Control. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023. [\[CrossRef\]](#)
15. Haruno, F.; Chonho, L.; Sakuei, O.; Hiromitsu, S. TEM-CD: A Template-based Evaluation Method for Controllable Difficulty in Stable Diffusion. *J. Inf. Process.* 2026, *in press*.
16. Hartwig, S.; Engel, D.; Sick, L.; Kniesel, H.; Payer, T.; Ropinski, T. Evaluating Text to Image Synthesis: Survey and Taxonomy of Image Quality Metrics. *arXiv* 2024, arXiv:2403.11821. [\[CrossRef\]](#)
17. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; Chen, X. Improved Techniques for Training GANs. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2234–2242.
18. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6629–6640. [\[CrossRef\]](#)
19. Koo, J.; Hernandez, J.; Haji-Ali, M.; Yang, Z.; Ordonez, V. Evaluating Text-to-Image Synthesis with a Conditional Fréchet chet Distance. *arXiv* 2025, arXiv:2503.21721.
20. Izzo, E.; Parolari, L.; Vezzaro, D.; Ballan, L. 7Bench: A Comprehensive Benchmark for Layout-guided Text-to-image Models. *arXiv* 2025, arXiv:2508.12919.
21. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Daejeon, Republic of Korea, 6–10 October 2015; pp. 234–241.
22. Wang, R.; Chen, Z.; Chen, C.; Ma, J.; Lu, H.; Lin, X. Compositional text-to-image synthesis with attention map control of diffusion models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–21 February 2024; pp. 5544–5552. [\[CrossRef\]](#)
23. Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph. (TOG)* 2023, 42, 1–10. [\[CrossRef\]](#)
24. Lee, J.; Lee, J.S.; Lee, J.H. CountCluster: Training-Free Object Quantity Guidance with Cross-Attention Map Clustering for Text-to-Image Generation. *arXiv* 2025, arXiv:2508.10710.
25. Wooyeolbaek, M. Attention-Map-Diffusers: Cross Attention Map Tools for Hugging Face Diffusers. 2024. Available online: <https://github.com/wooyeolBaek/attention-map-diffusers> (accessed on 24 September 2025).
26. Labs, B. Flux-Schnell. 2023. Available online: <https://huggingface.co/black-forest-labs/FLUX.1-schnell> (accessed on 24 September 2025).
27. Danish, S.; Sadeghi-Niaraki, A.; Khan, S.U.; Dang, L.M.; Tightiz, L.; Moon, H. A comprehensive survey of Vision–Language Models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Inf. Fusion* 2026, 126, 103623. [\[CrossRef\]](#)
28. Chung, J.; Lim, S.; Lee, S.; Yu, Y. MASS: Overcoming Language Bias in Image-Text Matching. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 2591–2599.
29. McRoy, S.W.; Channarukul, S.; Ali, S.S. YAG: A template-based generator for real-time systems. In Proceedings of the INLG’2000 Proceedings of the First International Conference on Natural Language Generation, Mitzpe Ramon, Israel, 12–16 June 2000; pp. 264–267.
30. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
31. Spiewak, G. Practical English Usage, 4th edn, fully revised. *ELT J.* 2018, 72, 448–451. [\[CrossRef\]](#)
32. Thomson, A.J.; Martinet, A.V. *A Practical English Grammar*; Oxford University Press: Oxford, UK, 1986.
33. Oxford University Press. Oxford 3000 and 5000 Word Lists. 2020. Available online: <https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000> (accessed on 24 September 2025).
34. Dinh, T.M.; Nguyen, R.; Hua, B.S. Tise: Bag of metrics for text-to-image synthesis evaluation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 594–609. [\[CrossRef\]](#)
35. Wang, Z.; Berman, M.; Rannen-Triki, A.; Torr, P.; Tuia, D.; Tuytelaars, T.; Gool, L.V.; Yu, J.; Blaschko, M. Revisiting evaluation metrics for semantic segmentation: Optimization and evaluation of fine-grained intersection over union. *Adv. Neural Inf. Process. Syst.* 2023, 36, 60144–60225.
36. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv* 2023, arXiv:2312.11805. [\[CrossRef\]](#)

37. Landis, J. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)] [[PubMed](#)]
38. Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; Levy, O. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 36652–36663.
39. Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; Dong, Y. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 15903–15935.
40. Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; Li, H. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv* **2023**, arXiv:2306.09341.
41. Yellapragada, S.; Graikos, A.; Prasanna, P.; Kurc, T.; Saltz, J.; Samaras, D. Pathldm: Text conditioned latent diffusion model for histopathology. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 5182–5191. [[CrossRef](#)]

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.