

Article

Improving Transferability of Adversarial Attacks via Maximization and Targeting from Image to Video Quality Assessment

Georgii Gotin ^{1,2,*} , Ekaterina Shumitskaya ^{1,3} , Dmitriy Vatolin ^{1,2,3}  and Anastasia Antsiferova ^{1,3} 

¹ AI Center, Lomonosov Moscow State University, Lomonosovsky ave 27b1, Moscow 119991, Russia; ekaterina.shumitskaya@graphics.cs.msu.ru (E.S.); dmitriy@graphics.cs.msu.ru (D.V.); aantsiferova@graphics.cs.msu.ru (A.A.)

² Faculty CMC, Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow 119991, Russia

³ Institute for Artificial Intelligence, Lomonosov Moscow State University, Lomonosovsky ave 27b1, Moscow 119991, Russia

* Correspondence: georgii.gotin@graphics.cs.msu.ru

Abstract

This paper proposes a novel method for transferable adversarial attacks from Image Quality Assessment (IQA) to Video Quality Assessment (VQA) models. Attacking modern VQA models is challenging due to their high complexity and the temporal nature of video content. Since IQA and VQA models share similar low- and mid-level feature representations, and IQA models are substantially cheaper and faster to run, we leverage them as surrogates to generate transferable adversarial perturbations. Our method, **MaxT-I2VQA** jointly **Maximizes** IQA scores and **Targets** IQA feature activations to improve transferability from IQA to **VQA** models. We first analyze the correlation between IQA and VQA internal features and use these insights to design a feature-targeting loss. We evaluate MaxT-I2VQA by transferring attacks from four state-of-the-art IQA models to four recent VQA models and compare against three competitive baselines. Compared to prior methods, MaxT-I2VQA increases the transferability of an attack success rate by 7.9% and reduces per-example attack runtime by 8 times. Our experiments confirm that IQA and VQA feature spaces are sufficiently aligned to enable effective cross-task transfer.

Keywords: video quality assessment; video quality metric; adversarial attack; cross-modal; transferability

1. Introduction

Automated Image Quality Assessment (IQA) and Video Quality Assessment (VQA) models are now integral components of many media processing pipelines: they are used by broadcasters, streaming services, video-hosting platforms, and content-distribution networks to monitor and optimize both the technical fidelity and the perceived aesthetics of visual content. Given the growing role of these models in encoding decisions, quality control, and moderation, their reliability under distortions, such as compression artifacts or intentional noise, becomes critical. A failure to robustly estimate perceived quality can lead to suboptimal user experience, incorrect quality-driven decisions, or exploitation by malicious actors.

Adversarial attacks expose vulnerabilities in IQA and VQA models by intentionally perturbing input images and videos to cause inaccurate quality estimation [1,2]. In the IQA/VQA context, adversarial manipulations can be employed to artificially raise predicted quality scores, thereby hiding visual defects and compromising automated quality



Academic Editors: Francesco Archetti and Fabrizio Messina

Received: 28 November 2025

Revised: 7 January 2026

Accepted: 19 January 2026

Published: 5 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.

monitoring [3,4]. Adversarial attacks are commonly categorized as white-box, where the attacker has full access to the model architecture and gradients, and black-box, where access is limited to model outputs. While white-box techniques achieve stronger attacks by leveraging gradient information, full access to production models is rarely available in realistic scenarios. Black-box attacks, which must reconstruct gradients or rely on heuristic methods, therefore present a more realistic threat model.

A particularly practical subclass of black-box attacks is transferable attacks, which are adversarial examples generated using surrogate models via white-box attacks that transfer to unseen target models without any queries to them. Recent studies have demonstrated high transferability across architectures and tasks, indicating that surrogate-based strategies can be highly effective even when the target model is deployed and inaccessible to the attacker [5–8]. Transferability between different perceptual assessment tasks, e.g., from IQA to VQA, is especially interesting: if perturbations crafted to fool IQA models can also manipulate VQA predictions, attackers could leverage widely available IQA models as proxies to influence more complex, less-accessible VQA systems.

Prior work introduced IC2VQA, a transferable cross-modal attack that leverages IQA surrogate models to influence VQA predictions [9]. That study demonstrated that perturbations optimized for IQA networks can significantly alter the correlation statistics of VQA outputs. However, several essential limitations remained: (1) the attack was not targeted, i.e., it did not control the direction of change (increase versus decrease) in the VQA score; (2) the study did not explicitly enforce or evaluate layer-level consistency between surrogate and target representations, which can be crucial for transferability; and (3) evaluations did not include state-of-the-art industry VQA models such as DOVER and VMAF, leaving open the question of practical impact against deployed systems.

This paper addresses these limitations by developing a targeted, transferable cross-modal attack that not only perturbs VQA predictions but also specifically increases quality scores. Our approach improves cross-model transferability by aligning explicit representations at intermediate layers of surrogate IQA networks and by incorporating temporal consistency and perceptual constraints to preserve visual plausibility in video sequences. We also evaluate the proposed method against modern VQA models, including DOVER and VMAF, and across multiple model architectures and publicly available datasets.

Our main contributions are:

- A targeted cross-modal transfer attack (from IQA to VQA) that is explicitly optimized to perform a targeted attack (increase VQA scores) while remaining imperceptible or minimally perceptible. Code is available at: <https://github.com/GeorgeGotin/MaxT-I2VQA> (accessed on 18 January 2026).
- A layer-consistency framework that aligns intermediate representations of surrogate IQA models with target VQA features to improve transferability.
- Temporal- and perceptual-regularization terms designed for video data to maintain frame-to-frame coherence.
- Extensive empirical evaluation against state-of-the-art VQA models (including DOVER and VMAF) and a range of deep-learning VQA models, with ablation studies that quantify the effect of each design choice.

2. Related Works

Image and Video Quality Assessment (IQA/VQA) models can be broadly categorized into full-reference (FR) and no-reference (NR) approaches. FR models such as LPIPS [10], VMAF [11], and PSNR compute a quality score for a distorted image or video relative to a pristine reference. They are typically used in scenarios where a reference is available

and minimal degradation is desired, such as compression optimization or transmission quality monitoring.

However, in many real-world applications, a reference signal is unavailable or impractical to obtain. In these cases, NR models are required to estimate quality directly from a single image or video. Examples include IQA models such as PaQ-2-PiQ [12], NIMA [13], SPAQ [14], DBCNN [15], and Linearity [16] as well as VQA models like VSFA [17], DOVER [18], MDTVSFA [19], and TiVQA [20]. NR models are especially relevant in settings such as recommendation systems, camera control and tuning, or large-scale quality monitoring, where only the test content is available.

The notion of “quality” in IQA/VQA handles multiple aspects. In many cases, models aim to capture aesthetic quality, which is a human-centric perception of visual appeal determined by attributes such as colorfulness, contrast, focus, composition, and temporal consistency in videos. In other contexts, technical quality is the focus, emphasizing properties that facilitate computational analysis or the reliable operation of downstream systems; for example, ensuring sufficient sharpness and clarity in license plate recognition or medical imaging.

Attacks on both image and video quality assessment metrics can be divided into two sets: white-box attacks, gray-box attacks, and black-box attacks.

2.1. White-Box Attacks

White-box attacks, such as FGSM [21], PGD [22], Ti-Patch [23], and IOI [24], generate adversarial perturbations by directly optimizing loss functions using model gradients, either for each image or video, or in the case of universal attacks (e.g., UAP [3]) by producing a single perturbation that generalizes across multiple inputs. These attacks are typically highly effective against the models they are crafted for yet still suffer from poor transferability, as shown by Zhang et al. [25] and Liu et al. [26]. Among the methods mentioned above, PGD remains the most relevant baseline for studying transferability, as IOI and UAP are inherently less transferable by design, and FGSM typically yields weaker perturbations than iterative methods such as PGD.

2.2. Gray-Box Attacks

Gray-box attacks operate under partial knowledge of the target model, typically with access only to its output scores and without access to its internal parameters or gradients. Consequently, gradients cannot be computed directly, and the attacker must approximate them through repeated model queries. Examples include **AttackVQA** [27] and **Square Attack** [28], both of which are query-based methods. By iteratively perturbing inputs and observing changes in the model’s output, these attacks can estimate gradient directions that closely approximate the true ones, achieving strong attack performance even without direct access to gradients. However, this procedure is inherently time-consuming, as it requires multiple forward passes—often hundreds or thousands of model evaluations per example. All gray-box methods are relevant to our study, as they address a similar challenge: effectively attacking a target VQA model with limited or no internal knowledge of its architecture while maintaining competitive performance.

2.3. Black-Box Attacks

Black-box attacks assume no access to the target model’s internals: no gradients, no parameters, and sometimes not even output scores. In practice, many black-box strategies therefore rely on transferability: adversarial perturbations are crafted using one or more surrogate models and then applied to the unseen target under the assumption that different models share similar feature geometry. This surrogate-based approach is widely used in image and video classification (e.g., I2V [5], PDCL-Attack [6], GAMA [7], and Cross-

Modal-Attack [8]), and several works have shown successful image-to-video transfer for classification tasks. Transfer-based black-box attacks typically optimize perturbations with respect to surrogate gradients, hoping those gradients approximate the target's. The main advantages are usability (the method requires no queries to the target) and speed. These attacks generally require fewer iterations and no online querying, but they suffer from reduced effectiveness compared with white-box or query-based methods because the surrogate and target may differ substantially.

Our work stands out from the existing research in this area because most successful transfer techniques were developed for classification, where the objective is to change a discrete label or margin; such objectives naturally produce strong, transferable gradient directions. In contrast, IQA/VQA are regression tasks that predict continuous quality scores. Regression targets have different loss landscapes and decision criteria, so classification-focused transfer methods do not translate directly. Adapting transfer attacks to manipulate continuous quality estimates, therefore, requires different surrogate objectives (e.g., maximizing predicted quality), representation alignment, and temporal/perceptual constraints to produce effective and visually plausible perturbations for VQA systems.

3. Materials and Methods

3.1. Problem Formulation

Let $x \in V \subset [0, 1]^{N \times C \times H \times W}$ be a video drawn from the set of videos V , and let $x_i \in I \subset [0, 1]^{C \times H \times W}$ be the i -th frame of that video ($i = 1, \dots, N$). Here, N is the number of frames, H and W are height and width, and C is the number of color channels. Let $M_V : V \rightarrow [0, 1]$ denote a VQA model and $M_I : I \rightarrow [0, 1]$ denote an IQA model. Higher model outputs correspond to higher perceived quality and are assumed to correlate with humans' mean opinion score (MOS).

Because our attack exploits internal feature representations, we treat each model as a composition of layers. An IQA model M_I can be defined as $M_I = h_1 \circ \dots \circ h_K$, where $h_j : P_{j-1} \rightarrow P_j$, $P_0 = I$, $P_K = [0, 1]$. For any k , we define the head and tail of M_I as

$$F_k = h_1 \circ \dots \circ h_k, \quad G_k = h_{k+1} \circ \dots \circ h_K, \quad (1)$$

so F_k maps inputs to the intermediate feature space P_k , and G_k maps P_k to the final output. Thus, $M_I = G_k \circ F_k$. In this study of black-box transferability, we assume that M_V is a black-box model: we can only query its outputs and have no access to its internal parameters or gradients. By contrast, we assume full white-box access to M_I .

The goal of an attack on M_V is to find an adversarial perturbation $\delta \in [0, 1]^{N \times H \times W \times C}$ that maximizes the VQA output:

$$\delta = \underset{\delta' \in [0, 1]^{N \times H \times W \times C}}{\operatorname{argmax}} \quad M_V(x + \delta'). \quad (2)$$

subject to an imperceptibility constraint $\|\delta\|_p \leq \epsilon$ for a chosen p -norm. Here, δ denotes the adversarial perturbation. This is a key distinction from previous methods, as the problem is more challenging than merely misleading the VQA model.

3.2. Motivation for the Proposed Method

The black-box setting motivates studying the transferability of scores and internal representations between M_V and M_I . Intuitively, higher-quality videos tend to contain higher-quality frames. To quantify this, we measured correlations between features of the VQA and IQA models, similarly to [5].

Concretely, we extracted features from all layers of several IQA models for frames from a set of videos, and likewise saved features from the VQA model for the same videos. Frames were split into 10% training and 90% testing. Using the training frames, we learned a linear mapping from IQA to VQA features then evaluated the correlation between the mapped IQA features and the VQA features on the test frames. The resulting layer-wise correlations are shown in Figure 1. Many layer pairs exhibit strong correlation (up to 88%), which supports the feasibility of transferring perturbations crafted using IQA models to the target VQA model.

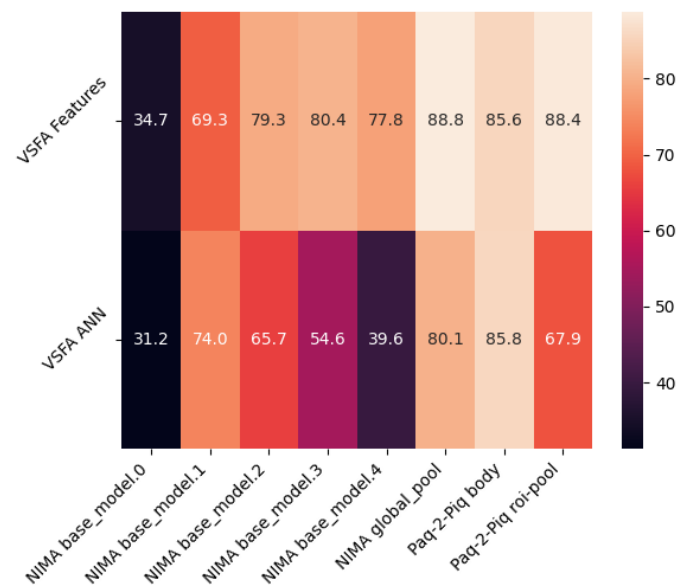


Figure 1. Layer-wise correlations between IQA and VQA features. For a set of videos, features were extracted from both IQA and VQA models. A linear mapping was learned on 10% of the frames; correlations reported are computed on the remaining 90%.

3.3. Proposed Method

The attack extends the IC2VQA [9] approach by retaining its core idea while adding several new losses and techniques to improve transferability from IQA to VQA (see the schematic in Figure 2).

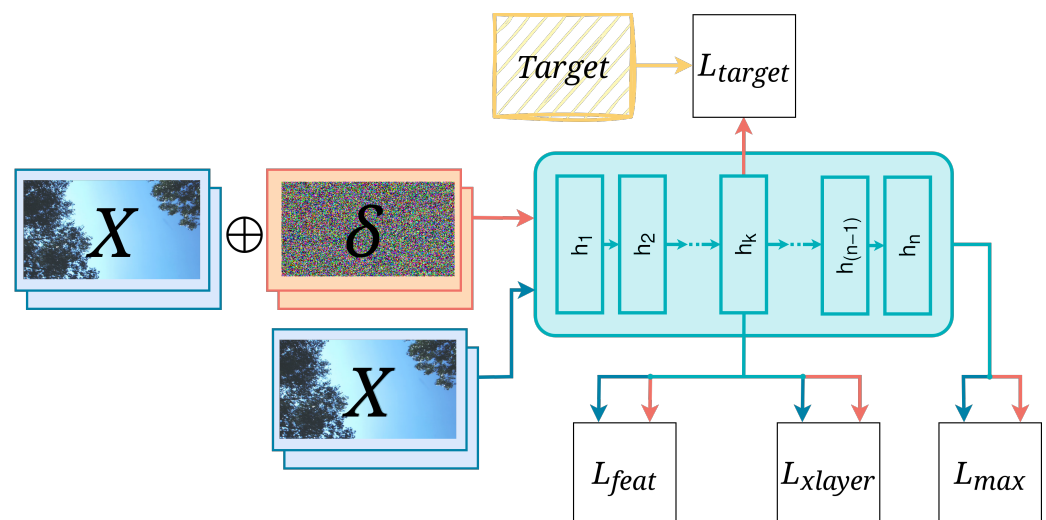


Figure 2. The scheme of the proposed method algorithm. Each frame of the video is perturbed by an adversarial attack and evaluated using an image quality metric. After extracting the features, they are used to optimize a loss of three components.

3.3.1. Cross-Layer Loss

Given the observed correlations between IQA and VQA feature spaces, we introduce a cross-layer loss that induces the perturbed frame features to differ from the original features in the IQA model's feature space. For the head F_k of M_I , the cross-layer loss is

$$\mathcal{L}_{xlayer} = \frac{1}{N} \sum_{i=1}^N \frac{F_k(x_i + \delta_i) \cdot F_k(x_i)}{\|F_k(x_i + \delta_i)\| \|F_k(x_i)\|}, \quad (3)$$

where x_i is the original i -th frame and δ_i is the i -th frame of the adversarial perturbation. Minimizing $\mathcal{L}_{xlayer}$ reduces the cosine similarity between original and perturbed frame features in the IQA feature space; due to the correlation between IQA and VQA feature spaces, this is expected to transfer and affect the VQA model's output.

3.3.2. Maximizing Loss

To directly push the IQA model's predicted quality upward (under the assumption that IQA and VQA scores are correlated), we add a loss that induces high IQA scores for perturbed frames:

$$\mathcal{L}_{max} = \frac{1}{N} \sum_{i=1}^N 1 - M_I(x_i + \delta_i). \quad (4)$$

Minimizing $\mathcal{L}_{max}$ increases the IQA model's outputs on perturbed frames, which should help raise the VQA output via cross-model correlations.

3.3.3. Target Loss

Building on the idea that driving intermediate activations toward those associated with a desired class can force a classifier to predict that class, we adapt this concept to quality prediction. We define a targeted cross-layer loss that pushes perturbed features toward a chosen target activation p_k in layer k :

$$\mathcal{L}_{target} = \frac{1}{N} \sum_{i=1}^N 1 - \frac{F_k(x_i + \delta_i) \cdot p_k}{\|F_k(x_i + \delta_i)\| \|p_k\|}, \quad (5)$$

where p_k is a target vector in the feature space P_k (for example, the layer output corresponding to a high-quality example). Minimizing $\mathcal{L}_{target}$ aligns perturbed features with p_k , and assuming that p_k corresponds to higher perceived quality, it should increase the VQA model's score.

3.3.4. Temporal and Feature Losses

Because video quality depends on temporal consistency and perturbations should remain imperceptible over time, we add temporal regularization losses that discourage the attack from introducing temporal artifacts (see the schematic in Figure 3).

Temporal smoothness on perturbations:

$$\mathcal{L}_{temp} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|\delta_{i+1} - \delta_i\|_2. \quad (6)$$

Feature-temporal consistency using IQA features of neighboring frames (cosine similarity):

$$\mathcal{L}_{feat} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{F_k(x_i) \cdot F_k(x_{i+1})}{\|F_k(x_i)\| \|F_k(x_{i+1})\|}. \quad (7)$$

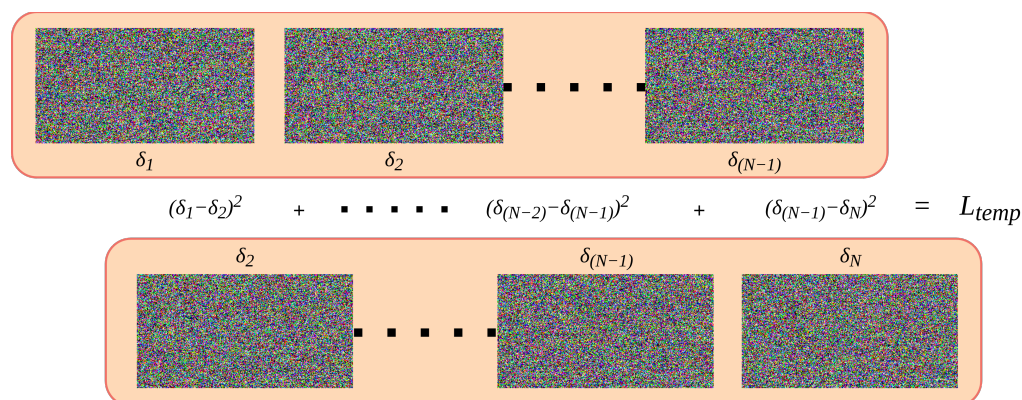


Figure 3. The algorithm of the temporal loss computing.

3.3.5. Summarizing Loss

All loss terms are combined linearly with coefficients α_j :

$$\mathcal{L} = \alpha_1 \mathcal{L}_{xlayer} + \alpha_2 \mathcal{L}_{max} + \alpha_3 \mathcal{L}_{target} + \alpha_4 \mathcal{L}_{temp} + \alpha_5 \mathcal{L}_{feat}. \quad (8)$$

where some α_j may be set to zero to disable particular terms. For stable optimization, we normalize losses so that their means are of similar magnitude, pushing them toward 1 during training.

3.3.6. Algorithm

Training on multiple IQA models can improve transferability, as an ensemble can capture richer, more general features. Our algorithm iterates over available IQA models and updates the perturbation δ using an optimizer (ADAM). At each iteration and for each IQA model, we compute the applicable loss terms, form the combined loss $\mathcal{L}$, take an optimizer step (with step size ϵ/I where I is the number of iterations), and clip δ to satisfy the budget $\|\delta\| \leq \epsilon$. The pseudo-code of the algorithm is presented in Algorithm 1.

Algorithm 1: Algorithm of the consistent attack with multiple IQA models

Data: original video $x \in [0, 1]^{N \times C \times H \times W}$, L IQA models
 $M_I^{(1)}, \dots, M_I^{(L)} : [0, 1]^{C \times H \times W} \rightarrow [0, 1]$, $k_1, \dots, k_L$ — numbers of the layer for each model, $p_{k_1}, \dots, p_{k_L}$ — targets for each layer, $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ — initial coefficients for each part of loss, perturbation budget ϵ , number of iterations I
Result: $\delta \in [0, 1]^{N \times C \times H \times W}$, s.t. $\|\delta\| \leq \epsilon$
 $\delta = (1/255)^{N \times C \times H \times W}$;
for i from 1 to I **do**
 for f from 1 to L **do**
 Calculate $\mathcal{L}_{xlayer}$ as in (3), $\mathcal{L}_{target}$ as in (5), $\mathcal{L}_{feat}$ as in (7) and for $M_I^{(f)}$;
 Calculate $\mathcal{L}_{temp}$ as in Equation (6)
 Calculate $\mathcal{L}$ as in Equation (8);
 $\delta \leftarrow \text{ADAM}(\epsilon/I, \mathcal{L})$;
 $\delta \leftarrow \text{clip}_\epsilon(\delta)$;
 Update $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$;
 end
end

Coefficients α_j are initialized to 1 for enabled losses and 0 for disabled ones; coefficients initialized to zero remain zero (so we can turn loss terms on or off). We additionally apply constraints so that the mean magnitudes of active coefficients remain near 1.

3.3.7. Target Generation

The targeted loss uses a precomputed target vector p_k for each chosen layer k . To compute p_k , we assume high-quality images form a cluster in the layer's feature space. Given a dataset of D images y_i , we run each image through the IQA model, obtaining both the model score $M_I(y_i)$ and the layer output $F_k(y_i)$. We then form a score-weighted average:

$$p_k = \frac{1}{D} \sum_{i=1}^D M_I(y_i)^\beta F_k(y_i). \quad (9)$$

where $\beta \geq 0$ is a power parameter that adjusts the influence of high-scoring images, with values closer to 1 giving stronger weighting to high scores (see the schematic in Figure 4). If a model's layer requires a specific input size, we resize images to the video frame size using bicubic interpolation; otherwise, we keep the original image size.

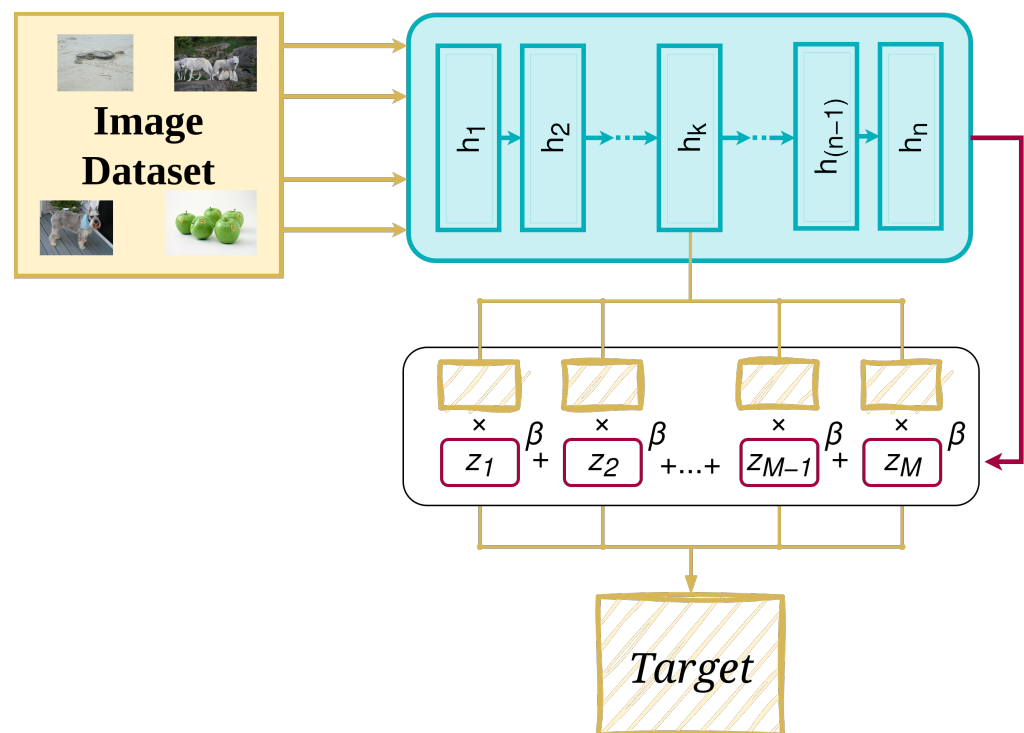


Figure 4. The scheme of feature target generating. For each image from the Image Dataset, the k -th layer output and model's output score z_i are computed. All results are summarized with weights z_i^β , where β is a parameter to form the target.

When limited black-box access to the VQA model is available, we also implemented a variant that uses a single call to M_V on each image treated as a single-frame video:

$$p_k = \frac{1}{D} \sum_{i=1}^D M_V([y_i])^\beta F_k(y_i). \quad (10)$$

3.3.8. Datasets

We constructed the target feature vectors from the KADID-10k image database (Kadid-10k) [29]. For each image, we computed IQA model scores and layer-wise feature maps and combined them as described to form p_k .

For evaluation, we used a subset of the Derf's dataset [30] comprising 10 videos, down-scaled from 1080p to 540p and trimmed to 75 frames each, and subset from CVQAD [31] comprising 54 videos, down-scaled from 1080p to 540p and trimmed to 10 frames each.

The videos were chosen to cover a range of motion patterns (tripod shots, moving crowds, running water, etc.). Each video was split into frames, and the proposed method was used to compute adversarial perturbations for each frame subject to the temporal regularizers described above.

3.3.9. IQA Model

For ensembles of IQA models, we used NIMA with layers `classifier` and `global_pool`, PaQ-to-PiQ with layers `body` and `roi_pool`, SPAQ with layers `layer1`, `layer2`, `layer3`, `layer4`, DBCNN with layers `features1` and `features2` with sublayers, and LPIPS with layers after AlexNet body. As LPIPS is an FR model, both original and adversarial frames are evaluated during the attack. Also, cross-layer and target losses are computed each time, since the layers are computed independently for each image. Additionally, distillations from several models such as NIMA, PaQ-2-PiQ, and SPAQ to Res-Net architectures were used. More details can be observed in the Table 1.

Table 1. The layers of IQA models, which are used for attacks.

Image Metrics	Layer Name	Layer Number	Layer Size
NIMA [13]	<code>classifier</code>	1	10
NIMA [13]	<code>global_pool</code>	2	1536
PaQ-2-PiQ [12]	<code>body</code>	1	261,120
PaQ-2-PiQ [12]	<code>roi_pool</code>	1	2048
SPAQ [14]	<code>layer1</code>	1	8,294,400
SPAQ [14]	<code>layer2</code>	2	4,177,920
SPAQ [14]	<code>layer3</code>	3	2,088,960
SPAQ [14]	<code>layer4</code>	4	1,044,480
DBCNN [15]	<code>features1</code>	30	1,013,760
DBCNN [15]	<code>features2</code>	57	261,120
LPIPS [10]	<code>slice5</code>	5	483,328

Four VQA models chosen for comparison are VSFA [17], TiVQA [20], and DOVER [18], as NR models, and one FR model, VMAF [11].

4. Results

All results are reported as percent relative gain with respect to the original (clean) target-model score.

$$\#gain\ per\ video = \left(\frac{\#score\ of\ video\ metric\ for\ attacked\ video}{\#score\ of\ video\ metric\ for\ original\ video} - 1 \right) * 100\%. \quad (11)$$

For each surrogate–target model pair, we ran the attack across the whole test set while sweeping the attack hyperparameter grid, which contains the number of epochs with values 1, 2, 5, 10, 15, and 20 and attack’s budget with values 1/255, 10/255, and 20/255. For each hyperparameter group, we averaged the measured gain across all videos; the reported value for that surrogate–target pair is the maximum of those averages. In other words, for each pair of surrogate and target, we (1) run the attack with all hyperparameter combinations on each video, (2) compute per-video gains, (3) average gains across videos for each hyperparameter group, and (4) report the best (largest) group-average gain.

4.1. Proposed Attack Effectiveness Across VQA Models

We evaluated IQA and VQA models over our dataset. The attack produced mostly positive outcomes, with performance gains of up to 27.8%; however, some attacks led to negative performance. The results are presented in Table 2.

Table 2. Table with results for attacking a single model. For LPIPS and VMAF, the branch with the original image stays unchanged. For VMAF, frame-by-frame evaluation was used. All results are given in percent gain relative to the original target model value. **Bold** for the best and underline for the second best.

Surrogate Model	DOVER [18]	TiVQA [20]	VMAF [11]	VSFA [17]
DBCNN [15]	−0.3363	−1.5317	27.3175	<u>4.5133</u>
Distillation to Res-Net-185	0.2881	0.2562	27.0754	0.7805
Distillation to Res-Net-503	0.3647	<u>1.2991</u>	27.1413	0.2369
Distillation to Res-Net-505	0.2755	0.7004	<u>27.2831</u>	0.6839
LPIPS [10]	−1.0568	−1.7240	27.0031	0.1999
NIMA [13]	1.4124	−0.4844	26.319850	0.880125
PaQ-2-PiQ [12]	<u>0.9767</u>	1.1908	26.537208	1.549308
SPAQ [14]	−0.0955	9.9503	27.805523	11.001774

4.2. Comparison with Other Methods

As discussed in Section 2, we compare our proposed method to three representative baselines: PGD [22], Square Attack [28], and VQA Attack [27]. For the PGD method, we generate adversarial perturbations for the surrogate IQA model and then evaluate their transferability to the target VQA metric. For the Square Attack, we directly attack the target VQA model using single-frame videos, where each frame corresponds to an original video frame. Finally, for the VQA Attack, we perform an adversarial attack on the target VQA directly using the full video and evaluate performance on the same sequence.

As shown in Table 3, our proposed method consistently outperforms the baselines, achieving mostly positive relative gains. Specifically, MaxT-I2VQA improves transfer effectiveness by up to 7.9%, which is achieved in comparison with Square Attack for TiVQA. Comparison for DOVER is omitted, as gains of other methods are negative and not representative. The negative results for DOVER highlight great robustness of the model.

Table 3. Comparison of the proposed transferable attack to other methods. For MaxT-I2VQA and FGSM surrogate models are mentioned. For AttackVQA and Square, the target models are used as surrogate ones. For all methods, the same number of epochs and budget limits were used; the best for each surrogate–target model pair is shown. All results are reported as percent gains relative to the original target model values. **Bold** for the best and underline for the second best.

Attack	Surrogate Model	DOVER [18]	TiVQA [20]	VMAF [11]	VSFA [17]
PGD [22]	NIMA [13]	−0.3579	−2.1224	25.7025	−0.5611
	PaQ-2-PiQ [12]	−0.5675	−2.4812	25.7024	−0.3811
	SPAQ [14]	−0.6164	−0.5926	25.7023	1.1925
Square [28]	–	−0.6119	<u>2.0326</u>	25.7024	<u>5.0994</u>
AttackVQA [27]	–	−1.2741	0.1888	–	0.4392
MaxT-I2VQA	LPIPS [10]	−0.5114	−1.7240	27.0031	0.1999
	NIMA [13]	1.4124	−0.4844	27.1911	0.8801
	PaQ-2-PiQ [12]	<u>0.9767</u>	1.1908	<u>27.3980</u>	1.5493
	SPAQ [14]	−0.0955	9.9503	27.8055	11.0018

We also measured the average time per iteration for each attack on videos with 540p resolution and 75 frames. The results are as follows: PGD—2.5 s, Square Attack—43 s, AttackVQA—254 s, and MaxT-I2VQA—5 s per iteration. These results demonstrate that MaxT-I2VQA is significantly faster than prior black-box methods—by at least 8 times—while still maintaining superior transferability. PGD is faster, but as shown in our comparisons, it exhibits poor transferability, limiting its practical usefulness for cross-model attacks.

5. Discussion

In this section, unlike the previous one, results are reported for a fixed target model and a selected hyperparameter set, while varying the surrogate model as an additional factor to design our method and observe its behavior on different parameters.

5.1. Discussion on Maximization and Targeting

Before tuning our method's parameters, we ran an additional experiment to identify the best attack variant, see results in the Table 4. The results of our experiments show that the optimal choice depends on the model, and therefore, the final attack uses both modification variants, selecting the better one in each case.

Table 4. Table with comparison of maximization and targeting losses.

Video Metric Modification	DOVER	TiVQA	VMAF	VSFA
No	−0.8542	−2.4938	26.1541	0.5314
Maximize	1.4124	9.9503	27.8055	11.0018
Targeted	−0.2733	7.0959	27.6993	9.9487

5.2. Discussion on Layer's Position

Figure 5 shows the impact of the depth of the layer on the gain for different surrogate models. As shown, for most models, the layer depth does not affect the gain. Despite this, the SPAQ model demonstrates strong correlation, with the late-mid layers having a greater impact, increasing it up to 2 times for TiVQA.

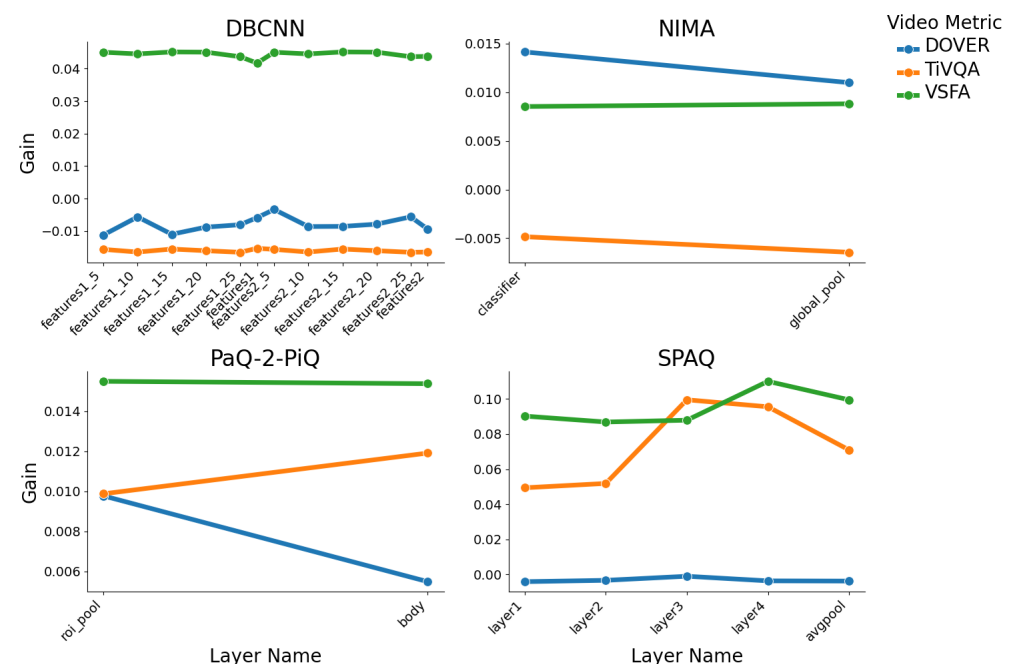


Figure 5. The study of gain across layers in different depths. Layers are sorted from left to right with the order of its appearance in the model.

5.3. Discussion on Attack Budget Limits

Our experiments show that, for transferable attacks, the relationship between attack success and budget is nonlinear, as shown in Table 5. For some model combinations, a smaller budget yields the best results, while others benefit from a larger budget. This can be explained by a different correlation of feature layers between IQA models and VQA

models. Supposedly, the method with low budget cannot generate attack with the best gain, but too big budget leads the method to optimize the attack for an IQA suppressing the correlation with a VQA model, so the gain decrease. Thus, the best budgets are individual for each VQA model and cannot be predicted easily.

Table 5. Table with comparison of attack with different budgets. Budgets are given for pixels range [0;1].

Video Metric Attack Strength	DOVER	TiVQA	VMAF	VSFA
1/255	0.5481	0.9929	27.8055	5.3477
10/255	−1.5144	9.9359	26.1541	11.0018
20/255	−0.2289	9.9503	25.2366	10.8381

5.4. Discussion on Steps Number

Our experiments show that the relationship between attack success and budget is nonlinear, as shown in Table 6. For each model, gains at low and high numbers of epochs are worse than at middle numbers, with different maximums.

Table 6. Table with comparison of attack with different numbers of iterations.

Video Metric Epoch Number	DOVER	TiVQA	VMAF	VSFA
1	0.3647	1.7387	27.7251	9.6060
2	0.5481	6.1937	27.8055	10.4075
5	0.3677	9.9503	27.7754	11.0018
10	−1.6009	0.8409	26.2598	1.5493
15	−1.6807	1.0621	26.2352	1.5252
20	−1.5180	−0.9078	25.7584	0.0349

5.5. Discussion on Ensembles

Our experiments demonstrate that ensembles can markedly improve attack performance, as shown in Table 7. With SPAQ, ensemble methods raise attack gain up to 13.9% on TiVQA, up to 9.5% on VSFA, and up to 27.7% on VMAF; the only exception is DOVER, where performance worsens.

Table 7. Table with results for ensembles of models. Both surrogate models are taken equally.

Surrogate Models	DOVER [18]	TiVQA [20]	VMAF [11]	VSFA [17]
DBCNN [15] with SPAQ [14]	−0.2091	3.2852	27.3463	8.9860
NIMA [13] with PaQ-2-PiQ [12]	−0.8228	0.7847	26.6606	0.3880
SPAQ [14] with Distillation to Res-Net-185	−0.0759	10.7966	27.7297	9.3510
SPAQ [14] with NIMA [13]	−0.3327	13.9750	27.7498	9.5300

5.6. Discussion on Parameters

All our experiments have shown that there is no linear dependency between the attack gain and any parameter. Each parameter–target model pair has its own dependency, which cannot be predicted, so during an attack, we need to run the method over all parameter–target pairs to achieve greater success.

6. Conclusions

This paper introduces MaxT-I2VQA, a black-box adversarial attack that is transferable from IQA to VQA models. Its primary goal is to generate imperceptible perturbations for videos using only IQA models as surrogates. Through extensive experiments, we

demonstrate that existing methods struggle to produce highly transferable perturbations, whereas our method achieves an attack success rate up to 7.9% higher. Moreover, MaxT-I2VQA is significantly more efficient, reducing attack time by at least 8 times compared to the next fastest method.

The proposed attack provides a powerful tool for experimentally assessing the vulnerability of VQA models. By releasing our method and code, we contribute to the development of safe and trustworthy video-quality assessment pipelines.

Author Contributions: Conceptualization, E.S. and A.A.; methodology, G.G., E.S. and A.A.; software, G.G.; validation, G.G. and E.S.; formal analysis, G.G., E.S. and A.A.; investigation, G.G. and E.S.; resources, G.G. and E.S.; data curation, E.S. and A.A.; writing—original draft preparation, G.G., E.S. and A.A.; writing—review and editing, E.S., D.V. and A.A.; visualization, G.G.; supervision, D.V. and A.A.; project administration, D.V. and A.A.; funding acquisition, D.V. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No 139-15-2025-012).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data can be found at <https://github.com/GeorgeGotin/MaxT-I2VQA> (accessed on 18 January 2026).

Acknowledgments: The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

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

References

1. Antsiferova, A.; Abud, K.; Gushchin, A.; Shumitskaya, E.; Lavrushkin, S.; Vatolin, D. Comparing the Robustness of Modern No-Reference Image- and Video-Quality Metrics to Adversarial Attacks. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 700–708.
2. Shumitskaya, E.; Antsiferova, A.; Vatolin, D. Fast Adversarial CNN-based Perturbation Attack on No-Reference Image- and Video-Quality Metrics. In Proceedings of the Tiny Papers ICLR, Kigali, Rwanda, 5 May 2023; Volume 38, pp. 700–708.
3. Shumitskaya, E.; Antsiferova, A.; Vatolin, D.S. Universal Perturbation Attack on Differentiable No-Reference Image- and Video-Quality Metrics. In Proceedings of the 33rd British Machine Vision Conference, London, UK, 21–24 November 2022.
4. Shumitskaya, T. Towards adversarial robustness verification of no-reference image- and video-quality metrics. *Comput. Vis. Image Underst.* **2024**, *240*, 103913. [[CrossRef](#)]
5. Wei, Z.; Chen, J.; Wu, Z.; Jiang, Y.-G. Cross-Modal Transferable Adversarial Attacks from Images to Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 15044–15053.
6. Yang, H.; Jeong, J.; Yoon, K.-J. Prompt-Driven Contrastive Learning for Transferable Adversarial Attacks. In Proceedings of the 18th European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 36–53.
7. Aich, A.; Ta, C.-K.; Gupta, A.; Song, C.; Krishnamurthy, S.; Asif, S.; Roy-Chowdhury, A. GAMA: Generative Adversarial Multi-Object Scene Attacks. In Proceedings of the 35th Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 36914–36930.
8. Wang, R.; Guo, Y.; Wang, Y. Global-Local Characteristic Excited Cross-Modal Attacks from Images to Videos. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 2635–2643. [[CrossRef](#)]
9. Gotin, G.; Shumitskaya, E.; Antsiferova, A.; Vatolin, D. Cross-Modal Transferable Image-to-Video Attack on Video Quality Metrics. In Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications—Volume 3: VISAPP, Porto, Portugal, 26–28 February 2025; pp. 880–888.
10. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.

11. VMAF: Perceptual Video Quality Assessment Based on Multi-Method Fusion. Available online: <https://github.com/Netflix/vmaf> (accessed on 1 October 2025).
12. Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; Bovik, A. From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 16–22 June 2020; pp. 3572–3582.
13. Talebi, H. NIMA: Neural Image Assessment. *IEEE Trans. Image Process.* **2018**, *27*, 3998–4011. [[CrossRef](#)] [[PubMed](#)]
14. Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; Wang, Z. Perceptual Quality Assessment of Smartphone Photography. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 16–18 June 2020; pp. 3677–3686.
15. Zhang, W.; Ma, K.; Yan, J.; Deng, D.; Wang, Z. Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 36–47. [[CrossRef](#)]
16. Li, D.; Jiang, T.; Jiang, M. Norm-in-Norm Loss with Faster Convergence and Better Performance for Image Quality Assessment. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 789–797.
17. Li, D.; Jiang, T.; Jiang, M. Quality Assessment of In-the-Wild Videos. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2351–2359.
18. Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.H.; Wang, A.; Sun, W.S.; Yan, Q.; Lin, W. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 20087–20097.
19. Li, D. Unified Quality Assessment of in-the-Wild Videos with Mixed Datasets Training. *Int. J. Comput. Vis.* **2021**, *129*, 1238–1257. [[CrossRef](#)]
20. Zhang, A.-X.; Wang, Y.-G. Texture Information Boosts Video Quality Assessment. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 2050–2054.
21. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
22. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
23. Leonenkova, V.; Shumitskaya, E.; Antsiferova, A.; Vatolin, D. Ti-Patch: Tiled Physical Adversarial Patch for no-reference video quality metrics. *arXiv* **2024**, arXiv:2404.09961.
24. Shumitskaya, E.; Antsiferova, A.; Vatolin, D.S. IOI: Invisible One-Iteration Adversarial Attack on No-Reference Image- and Video-Quality Metrics. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; pp. 45329–45352.
25. Zhang, W.; Li, D.; Min, X.; Zhai, G.; Guo, G.; Yang, X.; Ma, K. Perceptual Attacks of No-Reference Image Quality Models with Human-in-the-Loop. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 2916–2929.
26. Liu, Y.; Yang, C.; Li, D.; Ding, J.; Jiang, T. Defense against adversarial attacks on no-reference image quality models with gradient norm regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 25554–25563.
27. Zhang, A.-X.; Ran, Y.; Tang, W.; Wang, Y.-G. Vulnerabilities in Video Quality Assessment Models: The Challenge of Adversarial Attacks. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 51477–51490.
28. Andriushchenko, M.; Croce, F.; Flammarion, N.; Hein, M. Square Attack: A query-efficient black-box adversarial attack via random search. In Proceedings of the Computer Vision—ECCV 2020, SEC, Glasgow, UK, 23–28 August 2020; pp. 484–501.
29. Lin, H.; Hosu, V.; Saupe, D. KADID-10k: A Large-scale Artificially Distorted IQA Database. In Proceedings of the 11th International Conference on Quality of Multimedia Experience, Berlin, Germany, 5–7 June 2019; pp. 1–3.
30. Xiph.org Video Test Media. Available online: <https://media.xiph.org/video/derf/> (accessed on 1 October 2025).
31. Antsiferova, A.; Lavrushkin, S.; Smirnov, M.; Gushchin, A.; Vatolin, D.; Kulikov, D. Video compression dataset and benchmark of learning-based video-quality metrics. In Proceedings of the 35th Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 13814–13825.

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.