

Essence of image quality

Ruilin Li

*Program of Computer Science, City University of Hong Kong (Dongguan),
Guangdong, China, and*

Shan Huang

*School of International Politics and Communication,
Beijing Language and Culture University, Beijing, China*

APSIPA
Transactions on
Signal and
Information
Processing

707

Received 31 August 2025
Revised 22 October 2025
12 February 2026
Accepted 12 February 2026

Abstract

Perceptual image quality is a complex, multifaceted attribute influenced by factors such as sharpness, noise, contrast, and semantic content. Traditional image quality assessment methods measure overall quality but lack interpretable decompositions to isolate specific perceptual factors. In this work, we leverage the generative capacity of text-to-image diffusion models, which can be leveraged through textual inversion to decompose image quality into disentangled embeddings. By structuring these embeddings according to perceptual quality metrics such as Mean Opinion Score, our framework enables decomposing image quality while preserving semantic content. Moreover, this decomposition provides a principled way to investigate the underlying factors and essence of perceptual image quality, including the role of semantic content. To our knowledge, this is the first framework that decomposes image quality using the generative power of diffusion models.

Keywords Image quality assessment, Diffusion model, Textual inversion

Paper type Research paper

1. Introduction

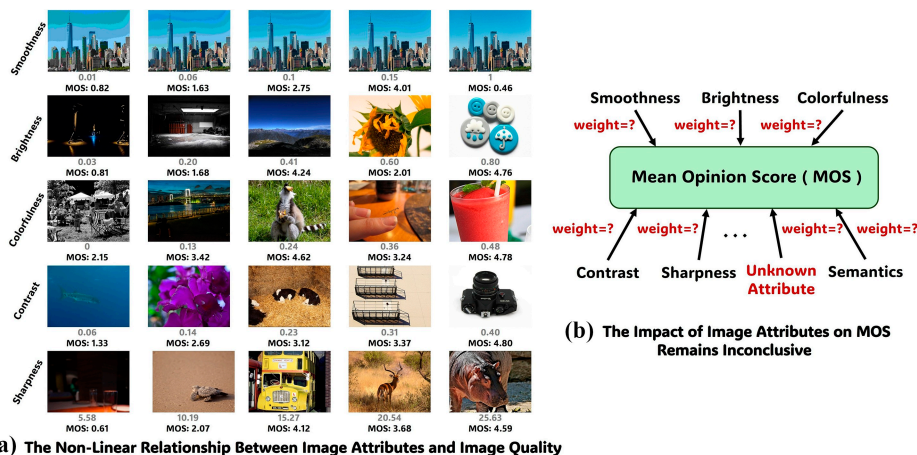
Perceptual image quality has long been a central topic in computer vision and image processing. The field of image quality assessment (IQA) has made significant progress in measuring quality, both with reference (FR-IQA) and without reference (NR-IQA) (Mittal *et al.*, 2012a; Sheikh and Bovik, 2006; Wang *et al.*, 2004). Deep neural network-based IQA methods (Bosse *et al.*, 2017; Kang *et al.*, 2014; Ke *et al.*, 2021; Zhang *et al.*, 2018) have achieved high correlation with human subjective ratings, such as Mean Opinion Score (MOS). However, while IQA excels at quantifying overall image quality, it lacks mechanisms to decompose quality into contributing factors. This decomposition is inherently challenging because perceptual quality is subjective, context-dependent and entangled with semantic content – altering one aspect of an image (e.g. sharpness or noise) can unintentionally affect perceived content or style.

As illustrated in Figure 1(a), the relationship between individual image attributes (e.g. sharpness, noise, contrast) and perceptual quality is highly nonlinear, making the impact of each attribute on MOS inconclusive. Moreover, as shown in Figure 1(b), interactions among multiple attributes further complicate their joint influence on perceived quality. Together,



© Ruilin Li and Shan Huang. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licences/by/4.0/>

APSIPA Transactions on Signal
and Information Processing
Vol. 15 No. 1, 2026
pp. 707-721
Emerald Publishing Limited
2048-7703
DOI 10.1108/ATSP-08-2025-0087



(a) The Non-Linear Relationship Between Image Attributes and Image Quality

Figure 1. Challenge of image quality decomposition: (a) MOS does not vary linearly with changes in individual image attributes; (b) the impact of single or multiple image attributes contributing on MOS remains unknown

these observations indicate that existing methods (Blau and Michaeli, 2018; Ke *et al.*, 2021; Mittal *et al.*, 2012a; Sheikh and Bovik, 2006; Wang *et al.*, 2004; Zhang *et al.*, 2018) – which typically measure only overall quality – cannot reveal which image attributes primarily drive perceptual judgments or how semantic content interacts with them. This underscores the difficulty of uncovering the essence of image quality.

Recently, text-to-image diffusion models (Ho *et al.*, 2020; Rombach *et al.*, 2022; Song *et al.*, 2020) such as stable diffusion (SD) (Rombach *et al.*, 2022) have demonstrated remarkable ability to generate high-fidelity, diverse and semantically coherent images. Beyond their generative capacity, these models support textual inversion (Gal *et al.*, 2022; Ruiz *et al.*, 2023), allowing optimization of a new token to represent a specific concept in embedding space. This capability opens the door to learning disentangled embeddings that capture perceptual quality independently of semantics. By leveraging the frozen backbone of SD and its text-conditioning mechanism, we can build an embedding space that explicitly separates semantic content from image quality.

In this work, we propose a novel framework for image quality decomposition, bridging IQA and generative modeling. Specifically, our approach first extracts semantic-preserving embeddings via textual inversion on SD, then learns a set of quality embeddings according to MOS differences. This enables not only controllable image synthesis but also a systematic study of the underlying essence of image quality, by disentangling perceptual factors from semantic content. To the best of our knowledge, this is the first work that systematically disentangles and controls image quality within powerful diffusion-based generative models.

2. Related work

2.1 Diffusion-based inversion

Diffusion models have demonstrated impressive capabilities in image generation, yet controlling the generation process requires effective inversion methods. A naive approach to inversion is to add noise to an image and then denoise it through the network, but this often leads to significant changes in image content. To address this (Choi *et al.*, 2021), conditions

the denoising process on noised low-pass filter data from the target image, while Denoising Diffusion Implicit Mode (DDIM)-based methods (Dhariwal and Nichol, 2021) enable closed-form inversion to extract latent noise maps capable of reproducing a given real image. These techniques have been extended to enable cross-image interpolations and semantic editing, often using CLIP-based latent vectors for conditioning (Ramesh *et al.*, 2021). However, such approaches are typically limited when generating novel conditions for a subject while maintaining fidelity.

In text-to-image diffusion models, inversion can be performed in two main latent spaces: the Image-based Noise Map (INM) space and the Textual Word Embedding (TWE) space (Dong *et al.*, 2025; Gal *et al.*, 2022; Kumari *et al.*, 2023; Li *et al.*, 2015; Qiu, 2016; Qiu *et al.*, 2015, 2019b; Ruiz *et al.*, 2023; Wang *et al.*, 2018). INM-based methods, such as DDIM (Song *et al.*, 2020) and Null Text (Mokady *et al.*, 2023), recover the initial noise to faithfully reconstruct the image but generally suffer from limited editing capabilities. In contrast, TWE-based methods, including Textual Inversion (Gal *et al.*, 2022) and DreamArtist (Dong *et al.*, 2025), optimize the embedding of new pseudo-words to represent user-provided concepts, allowing for more flexible and intuitive textual editing. DreamBooth (Ruiz *et al.*, 2023) further fine-tunes the entire model to learn high-fidelity concepts with unique identifiers, while Custom Diffusion (Kumari *et al.*, 2023) improves efficiency by updating only key and value mappings in cross-attention layers. Despite these advances, learning a single new concept can still take several minutes, limiting scalability.

Textual inversion methods leverage the semantic richness of pre-trained text-to-image models to efficiently represent new concepts (Gal *et al.*, 2022; Qiu *et al.*, 2019a, 2020, 2021; Sun *et al.*, 2021; Xu *et al.*, 2022; Yang *et al.*, 2020a, 2020b). By introducing new pseudo-words into the text embedding space, these methods enable personalized image generation through standard text prompts. Gradient-based optimization (Cui *et al.*, 2025; Li *et al.*, 2025; Ren *et al.*, 2025a, 2025b; Ruder, 2016; Tang *et al.*, 2026; Xu *et al.*, 2024; Yu *et al.*, 2024; Zhang *et al.*, 2025a, 2025b) with calibrated learning rates ensures that the pseudo-word embeddings faithfully capture the target concept without catastrophic forgetting. This framework allows inversion techniques to not only reconstruct images but also provide a decomposition of visual concepts semantically. These capabilities of diffusion models provide a natural foundation for learning disentangled embeddings, making image quality decomposition feasible.

2.2 Image quality assessment

Extensive subjective experiments have been conducted to understand human perception of image quality, leading to the establishment of numerous image quality assessment (IQA) data sets. For instance, the LIVE database (Sheikh *et al.*, 2006) collected subjective ratings for images degraded by JPEG, JPEG2000, white noise, Gaussian blur and simulated Rayleigh fading channel distortions under strictly controlled laboratory conditions. In addition to these controlled data sets, several large-scale IQA data sets have been created via crowdsourcing, such as TID2008 (Ponomarenko *et al.*, 2009), KonIQ-10k (Hosu *et al.*, 2020) and LIVE In the Wild (Ghadiyaram and Bovik, 2015). TID2008 contains 1,700 images covering a wide range of distortions, including noise, blur, compression artifacts, transmission errors and variations in brightness and contrast. KonIQ-10k was designed to capture real-world conditions, collecting 1.2 million subjective ratings from 1,459 participants across 10,073 images. The LIVE In the Wild data set comprises 1,162 images captured with diverse mobile devices, aiming to evaluate authentic distortions arising during capture, processing and storage. While databases such as TID2013 and KADID-10k focus on

artificially introduced distortions, LIVE In the Wild emphasizes the evaluation of naturally occurring degradations.

Building upon these subjective studies, various computational IQA methods have been developed for quality prediction. Depending on the availability of reference images, these models are typically categorized as full-reference (FR), reduced-reference (RR) or no-reference (NR) IQA. Furthermore, IQA approaches can be divided into traditional methods relying on handcrafted features and deep learning-based methods that automatically learn relevant representations. Classical FR-IQA techniques include structural similarity index measure (SSIM) (Wang *et al.*, 2004), MS-SSIM (Wang *et al.*, 2003) and VIF (Sheikh and Bovik, 2006), whereas deep learning-based FR-IQA methods include DR-IQA (Ghildyal and Liu, 2022) and IQT (Yin *et al.*, 2022). The proliferation of digital media has increased interest in NR-IQA, with traditional approaches often relying on natural scene statistics (Gao *et al.*, 2013; Mittal *et al.*, 2012a, 2012b; Moorthy and Bovik, 2010) or engineered features (Ma *et al.*, 2017; Ye and Doermann, 2012; Ye *et al.*, 2014, 2012).

Recently, deep learning has significantly advanced IQA performance. Convolutional neural network-based methods, such as WaDIQaM (Bosse *et al.*, 2017) and DB-CNN (Zhang *et al.*, 2020), have shown strong capability in extracting quality-relevant features. ResNet-based HyperIQA (Su *et al.*, 2020) distinguishes between low- and high-level features and leverages high-level information to refine low-level representations. MetaIQA (Zhu *et al.*, 2020) employs meta-learning to adapt networks to specific distortion types, enhancing prior knowledge utilization. Transformer-based approaches, including MANIQA (Yang *et al.*, 2022) and DEIQT (Qin *et al.*, 2023), exploit attention mechanisms for feature aggregation and efficient scoring, while StairIQA (Sun *et al.*, 2023) integrates hierarchical features with iterative mixed-database training for improved generalization. More recently, AGAIQA (Wang *et al.*, 2024) introduces adaptive graph attention and patch-wise hierarchical regression to further optimize feature exploitation and scoring precision.

However, despite these advances in perceptual quality assessment, no existing work has addressed the decomposition of image quality into factors and investigated whether semantics affect perceived quality.

3. Methodology

3.1 Overview

We propose a framework that aim to investigate how different quality factors affect image perception by generating images with varying quality while preserving semantic content (Figure 2).

3.2 Preliminary

3.2.1 Latent diffusion models. We build our approach upon latent diffusion models (LDMs) (Rombach *et al.*, 2022), a recently proposed class of denoising diffusion probabilistic models (DDPMs) (Ho *et al.*, 2020) that operate in the latent space of an autoencoder. LDMs comprise two main components. First, an autoencoder is pre-trained on a large image collection. An encoder E maps an image $x \in \mathcal{X}$ to a latent code $z = \mathcal{E}(x)$, regularized via either a KL-divergence loss or vector quantization (Van Den Oord *et al.*, 2017). The decoder D reconstructs images from latents such that $D(\mathcal{E}(x)) \approx x$.

The second component is a diffusion model trained to generate latents in the learned space. This model can be conditioned on class labels, segmentation masks, or outputs from a jointly trained text-embedding model. Let $c_\theta(y)$ denote a conditioning network that maps input y to a vector. The LDM loss is defined as:

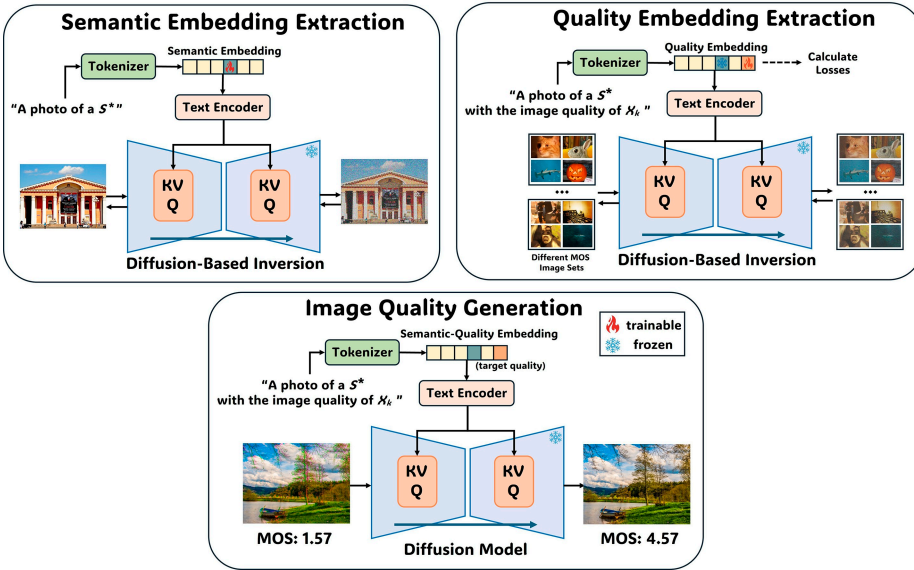


Figure 2. Illustration of our image quality decomposition framework. Using a frozen stable diffusion model, we first extract a semantic embedding via textual inversion. We then learn quality embeddings from different MOS image sets that are aligned with subjective scores and semantic content. During inference, the semantic embedding is combined with an interpolated quality embedding to generate images that maintain the original content while reflecting the target perceptual quality

$$L_{LDM} := \mathbb{E}_{z \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0, 1), t} \left[\|\epsilon - \epsilon_{\theta}(z_t, t, c_{\theta}(y))\|_2^2 \right] \quad (1)$$

where t is the diffusion timestep, z_t is the latent with noise added at step t , ϵ is the sampled noise and θ is the denoising network. Intuitively, the model learns to remove noise from latent representations. During training, both c_{θ} and θ are optimized jointly to minimize [equation \(1\)](#). At inference, a random noise tensor is sampled and iteratively denoised to obtain z_0 , which is then decoded into an image $x' = D(z_0)$. Besides, c_{θ} corresponds to a CLIP text encoder ([Radford et al., 2021](#)), with y as the text prompt.

3.2.2 Text embeddings. Text encoder models, such as CLIP, tokenize an input string and map each token to a learned embedding vector. In our approach, we target this embedding space for concept inversion. Specifically, we introduce a placeholder string S^* representing a novel concept and replace its token embedding with a learnable vector v^* . This allows us to compose new prompts containing the concept as with standard words.

3.2.3 Textual inversion. To learn v^* , we use a set of images depicting the target concept in varied poses and backgrounds. The concept embedding is optimized by minimizing the LDM loss over images from this set. Concept extraction is achieved using neutral context templates derived from CLIP ImageNet prompts ([Radford et al., 2021](#)), e.g. "A photo of S^* ". Specifically, the embedding inversion optimization is expressed as:

$$v_* = \arg \min_v \mathbb{E}_{z \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2 \right] \quad (2)$$

where the training procedure reuses the original LDM training scheme, keeping both c_θ and θ fixed. As this is a reconstruction task, the learned embedding captures fine-grained details unique to the concept.

3.3 Definition

The influence of semantic content and low-level factors on image quality remains largely unknown. For instance, images with little or overly simple semantic content often receive lower subjective quality scores, while low-level factors such as blur or pixel noise also typically lead to lower quality ratings. However, the precise effects of these factors on image quality, as well as their potential interactions, are not yet fully understood. Therefore, in our methodology, we first fix the semantic content and then explore the impact of varying low-level factors on image quality. This approach allows us to better isolate and analyze the specific contributions of low-level factors when the semantic content remains constant.

Our goal is to disentangle semantic content from perceptual image quality, and to enable direct generation of quality levels while preserving semantics. We build upon a pre-trained SD model, which remains entirely frozen during training. The key idea is to introduce learnable tokens through textual inversion: one token represents the semantic content of an image, while another set of tokens captures different levels of image quality. We describe the process in three stages.

3.4 Semantic embedding extraction

Given an image I_i with a subjective quality score m_i , we first isolate its semantic content. To do so, we define a textual prompt of the form a photo of S^* where S^* is a learnable placeholder token. Using textual inversion, we optimize the embedding v_i^* of S^* such that the frozen diffusion model can faithfully reconstruct I_i . Concretely, we minimize the standard denoising loss:

$$\mathcal{L}_{\text{TI}}^{(i)} = \mathbb{E}_{t, \epsilon} \|\epsilon - \epsilon_\phi(\alpha_t z_{0,i} + \sigma_t \epsilon, t, E_\theta(\text{"a photo of } S^*"; v_i^*))\|_2^2 \quad (3)$$

where $z_{0,i}$ is the VAE latent of I_i , and α_t, σ_t denote the noise schedule. After convergence, v_i^* provides a frozen semantic representation for image i , disentangled from its perceptual quality.

After obtaining v_i^* , we further exploit the key property of classifier-free guidance, which is highly influenced by the unconditional prediction branch. Instead of modifying the model weights or conditional embeddings, we adopt null-text optimization (Mokady et al., 2023), which refines only the unconditional embedding $\emptyset$. Specifically, for each input image, $\emptyset$ is initialized with the default null-text embedding and optimized while keeping the model parameters and conditional textual embedding fixed. This optimization yields a high-fidelity reconstruction of the input image while preserving editability through Prompt-to-Prompt (Hertz et al., 2022) by simply reusing the optimized unconditional embedding. Moreover, the optimized $\emptyset$ can be reused for multiple editing operations after a single inversion, providing efficiency and consistency across edits. Since null-text optimization is inherently less expressive than model fine-tuning, it is performed under the pivotal inversion framework to ensure efficient and stable convergence.

3.5 Quality embeddings extraction

Once the semantic token is fixed, we turn to modeling image quality. We discretize the MOS scores into K bins, each corresponding to a canonical quality level with center μ_k . For each bin, we introduce a quality token X_k with embedding u_k . To supervise these embeddings, we construct prompts of the form "a photo of S^* with the image quality of X_k " as P , which jointly condition on semantics and quality information.

3.5.1 Reconstruction consistency. For a sample I_i belonging to bin k_i , we optimize the diffusion loss:

$$\mathcal{L}_{\text{diff}}^{(i)} = \mathbb{E}_{t, \epsilon} \|\epsilon - \epsilon_\phi(\alpha_t z_{0,i} + \sigma_t \epsilon, t, E_\theta(P; [v_i^*, u_{k_i}]))\|_2^2 \quad (4)$$

This ensures that the model reconstructs I_i when conditioned on the correct semantic embedding v_i^* and its assigned quality embedding u_{k_i} .

3.5.2 Structuring quality spaces. Simply fitting each sample independently may not guarantee a coherent embedding space. We therefore impose a structure aligned with MOS scores. Specifically, we define a target similarity between quality embeddings and minimize as:

$$s_{k\ell} = \exp(-\alpha|\mu_k - \mu_\ell|), \quad (5)$$

$$\cos(u_k, u_\ell) = \frac{u_k^\top u_\ell}{\|u_k\| \|u_\ell\|}, \quad (6)$$

$$\mathcal{L}_{\text{cos}} = \sum_{k \neq \ell} \left(\cos(u_k, u_\ell) - s_{k\ell} \right)^2. \quad (7)$$

To further enforce relative ordering in the embedding space, we introduce a triplet margin loss based on MOS similarity. For each anchor embedding u_k , positive samples are drawn from MOS-similar bins and negative samples from MOS-dissimilar bins:

$$\mathcal{P}(k) = \{p | |\mu_k - \mu_p| \leq \delta\}, \quad \mathcal{N}(k) = \{n | |\mu_k - \mu_n| \geq \Delta\} \quad (8)$$

The triplet loss is then defined as:

$$\mathcal{L}_{\text{tri}} = \sum_k \mathbb{E}_{p \in \mathcal{P}(k), n \in \mathcal{N}(k)} [\cos(u_k, u_n) - \cos(u_k, u_p) + \gamma] \quad (9)$$

In our implementation, MOS scores are discretized into $K = 8$ equal-width bins, which is used to determine MOS-similar and MOS-dissimilar samples when constructing the positive and negative sets.

3.5.3 Image quality decomposition. To avoid leakage of semantic information into quality tokens, we encourage orthogonality between the semantic embedding v_i^* and all quality embeddings:

$$\mathcal{L}_\perp^{(i)} = \frac{1}{K} \sum_{k=1}^K \left(\frac{v_i^{*\top} u_k}{\|v_i^*\| \|u_k\|} \right)^2 \quad (10)$$

3.5.4 Overall losses. The full loss combines sample-wise reconstruction with global embedding structure:

$$\mathcal{L} = \sum_i \left(\mathcal{L}_{\text{diff}}^{(i)} + \lambda_{\perp} \mathcal{L}_{\perp}^{(i)} \right) + \lambda_{\text{cos}} \mathcal{L}_{\text{cos}} + \lambda_{\text{tri}} \mathcal{L}_{\text{tri}} \quad (11)$$

where $\lambda_{\perp}$, λ_{cos} , λ_{tri} are balancing coefficients. Only the semantic tokens v_i^* and the quality embeddings $\{u_k\}$ are optimized; the diffusion model itself remains frozen.

3.6 Image quality generation

At test time, given a target image I_{tar} , we first obtain its semantic embedding v_{tar}^* using the procedure in [equation \(3\)](#). The user specifies a desired MOS value m^{tar} , which we map to a continuous quality embedding via soft interpolation:

$$w_k(m^{\text{tar}}) = \frac{\exp(-\beta|m^{\text{tar}} - \mu_k|)}{\sum_{\ell=1}^K \exp(-\beta|m^{\text{tar}} - \mu_{\ell}|)}, \quad (12)$$

$$\tilde{u}(m^{\text{tar}}) = \sum_{k=1}^K w_k(m^{\text{tar}}) u_k. \quad (13)$$

The final conditioning is constructed by combining semantic and quality information. Concretely, we concatenate the semantic embedding v_{tar}^* with the quality embedding $\tilde{u}(m^{\text{tar}})$ and feed them into the text-conditioning encoder together with the prompt "a photo of v_{tar}^* with the image quality of $\tilde{u}(m^{\text{tar}})$ " to obtain the conditioning vector c_{tar} . We then use c_{tar} in the frozen diffusion model to generate $\hat{I} = \mathcal{D}(z_0)$, which preserves the semantic content of I_{tar} while matching the desired perceptual quality. By varying the target MOS value m^{tar} , the model produces images with smoothly controllable quality levels.

4. Experiments

4.1 Data sets

To evaluate the effectiveness of our image quality decomposition framework, we conduct experiments on widely used image quality assessment benchmarks, including LIVE IQA ([Sheikh et al., 2006](#)) and TID2013 ([Ponomarenko et al., 2015](#)).

LIVE IQA data set ([Sheikh et al., 2006](#)) is a widely used benchmark for evaluating image quality assessment algorithms. It contains 29 reference images and 779 distorted images generated from these references using five common distortion types: JPEG compression, JPEG2000 compression, Gaussian blur, white noise and fast fading distortions. Each distorted image is associated with a MOS obtained from subjective experiments, providing a reliable measure of perceived image quality. Due to its diverse distortion types and well-calibrated subjective scores, LIVE IQA is extensively used for benchmarking image quality assessment methods.

TID2013 data set ([Ponomarenko et al., 2015](#)) is a large-scale IQA benchmark consisting of 25 reference images and 3,000 distorted images. The distortions cover 24 different types, including additive noise, blur, compression artifacts, color distortions and spatial transformations, each applied at five severity levels. Similar to LIVE, TID2013 provides MOS values obtained through extensive subjective evaluation. Its large variety of distortion types and levels makes it particularly suitable for evaluating the robustness of image quality generation and enhancement algorithms under diverse degradation conditions.

4.2 Metrics

To quantitatively evaluate the performance of our image quality generation framework, we employ three widely used metrics: peak signal-to-noise ratio (PSNR), SSIM (Wang *et al.*, 2004) and Learned perceptual image patch similarity (LPIPS) (Zhang *et al.*, 2018). PSNR measures pixel-wise fidelity, SSIM evaluates perceptual similarity in terms of luminance, contrast, and structure, while LPIPS captures perceptual similarity using deep feature representations. Higher PSNR and SSIM values indicate better reconstruction quality, whereas lower LPIPS values indicate higher perceptual similarity.

4.3 Comparisons

We compare our method with three baseline approaches: Caption Prompt, Textual Inversion and DreamBooth.

Caption Prompt uses GPT-4 (Achiam *et al.*, 2023) to generate textual descriptions that specify the target image quality. These prompts are fed into the pre-trained text-to-image model to decompose image quality.

Textual Inversion (Gal *et al.*, 2022) learns new token embeddings for the target image quality using a small set of images. It can then generate new images reflecting the image quality with the learned embedding.

DreamBooth (Ruiz *et al.*, 2023) fine-tunes the text-to-image model on a few images of the target image quality, enabling high-fidelity generation with strong identity preservation.

As shown in Table 1, our method achieves the best overall performance on both LIVE IQA and TID2013 across all three evaluation metrics. In particular, compared to DreamBooth, which already provides strong reconstruction quality, our approach yields further improvements in PSNR, SSIM, and LPIPS, indicating better fidelity and perceptual similarity. While Textual Inversion also enhances performance compared to Caption Prompt, its gains are limited due to the restricted expressive capacity of token-level embeddings. Caption Prompt, which relies solely on GPT-4 generated text descriptions, performs the worst since the quality guidance in natural language is often ambiguous and insufficient to capture fine-grained distortions. These results demonstrate that our method more effectively leverages textual conditioning for image quality generation, leading to consistent improvements under diverse degradation conditions.

4.4 Ablation studies

As shown in Table 2, removing the quality embeddings leads to the most significant degradation, confirming that explicit quality tokens are essential for disentangling perceptual quality from semantic content. Without these embeddings, the model collapses to generic reconstructions with markedly lower PSNR and SSIM and higher LPIPS. Removing the cosine similarity loss also degrades performance, indicating that it plays a dominant role in structuring the quality embedding space to align with MOS-based perceptual distances and enabling smooth transitions across quality levels. In contrast, removing the triplet margin loss results in only a minor performance drop, as this component mainly refines class boundaries and enhances separability between distant quality levels rather than contributing substantially to reconstruction fidelity. Together, these results demonstrate that all three components are beneficial, with the cosine similarity loss and quality embeddings being the primary factors driving perceptual consistency and accuracy in our framework.

4.5 User studies

To further assess the effectiveness of our framework, we conducted a user study with 40 participants. Each participant was asked to complete two tasks. In the first task, they were

Table 1. Quantitative comparison of image quality generation methods on LIVE IQA (Sheikh *et al.*, 2006) and TID2013 (Ponomarenko *et al.*, 2015) data sets. Higher PSNR and SSIM indicate better reconstruction quality, while lower LPIPS indicates higher perceptual similarity

Method	LIVE IQA (Sheikh <i>et al.</i> , 2006)		TID2013 (Ponomarenko <i>et al.</i> , 2015)		LPIPS ↓
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	
Caption Prompt	22.8	0.841	21.9	0.832	0.369
Textual Inversion (Gal <i>et al.</i> , 2022)	23.9	0.867	22.8	0.857	0.348
DreamBooth (Ruiz <i>et al.</i> , 2023)	24.1	0.872	23.4	0.863	0.335
Ours	24.3	0.879	23.9	0.871	0.322

Table 2. Ablation study of our method on LIVE IQA (Sheikh *et al.*, 2006) and TID2013 (Ponomarenko *et al.*, 2015) data sets. Higher PSNR/SSIM and lower LPIPS are better

Variant	LIVE IQA (Sheikh <i>et al.</i> , 2006)			TID2013 (Ponomarenko <i>et al.</i> , 2015)		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o quality embeddings	22.8	0.835	0.362	22.4	0.828	0.381
w/o cosine loss	23.6	0.866	0.323	23.2	0.861	0.338
w/o triplet margin loss	24.0	0.874	0.313	23.7	0.867	0.328
Ours	24.3	0.879	0.305	23.9	0.871	0.322

presented with a reference image and four restored images produced by Caption Prompt, Textual Inversion, DreamBooth and our method. Participants were asked to select the restored image whose perceptual quality best matched the reference image’s MOS score. In the second task, participants compared image pairs generated by different methods and indicated which one appeared closer in perceptual quality to the reference.

As shown in Table 3, our method achieves the highest user preference in both tasks. Specifically, 38.2% of participants selected our results as most consistent with reference MOS scores in Task 1, and 34.2% preferred our method in direct pairwise comparisons. In contrast, DreamBooth ranked second but was consistently outperformed by our framework, while Caption Prompt and Textual Inversion lagged behind. These results corroborate the quantitative metrics and confirm that our approach provides perceptual advantages that are evident to human observers, consistent with prior findings on the importance of perceptual validation in image restoration (Sheikh *et al.*, 2006; Wang *et al.*, 2004; Zhang *et al.*, 2018).

5. Conclusion

In this work, we introduced a novel framework for decomposing and analyzing perceptual image quality using text-to-image diffusion models. By leveraging textual inversion in SD, our method learns semantic-preserving embeddings alongside disentangled quality embeddings structured by perceptual metrics such as MOS. This enables systematic analysis of how perceptual quality is formed while maintaining semantic consistency, offering new insights into the essence of human visual perception. To our knowledge, this is the first approach to explicitly separate semantic and low-level perceptual factors using generative diffusion models. Future work will focus on improving cross-data set generalization, addressing differences in distortion types, severity distributions and subjective scoring across data sets. We also plan to extend the framework toward multidimensional perceptual embeddings that jointly capture distinct visual factors (e.g. sharpness, noise, contrast) for more robust and transferable modeling.

Table 3. User preference results (%) across 40 participants. Higher values indicate stronger preference for the method

Method	Task 1: MOS alignment $\uparrow$	Task 2: Pairwise preference $\uparrow$
Caption Prompt	15.0	17.5
Textual Inversion (Gal <i>et al.</i> , 2022)	21.3	22.0
DreamBooth (Ruiz <i>et al.</i> , 2023)	25.5	26.3
Ours	38.2	34.2

Acknowledgements

This research project was supported by International Communication Research Center of China Federation of Radio and Television Associations (2024C007).

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023), "Gpt-4 technical report", *arXiv preprint arXiv:2303.08774*.
- Blau, Y. and Michaeli, T. (2018), "The perception-distortion tradeoff", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228-6237.
- Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T. and Samek, W. (2017), "Deep neural networks for no-reference and full-reference image quality assessment", *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, Vol. 15 No. 3, pp. 206-219.
- Choi, J., Kim, S., Jeong, Y., Gwon, Y. and Yoon, S. (2021), "ILVR: conditioning method for denoising diffusion probabilistic models", *arXiv preprint arXiv:2108.02938*.
- Cui, Y., Liu, L., Yu, B., Qiu, J., Zhang, X., Xiao, L., Liu, Y. and Chen, Q. (2025), "SPOGW: a score-based preference optimization method via group-wise comparison for workflows", *arXiv preprint arXiv:2510.04089*.
- Dhariwal, P. and Nichol, A. (2021), "Diffusion models beat GANS on image synthesis", *Advances in Neural Information Processing Systems*, Vol. 34, pp. 8780-8794.
- Dong, Z., Wei, P. and Lin, L. (2025), "DreamArtist: controllable one-shot text-to-image generation via positive-negative adapter", *International Journal of Computer Vision*, Vol. 133 No. 10, pp. 1-17.
- Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G. and Cohen-Or, D. (2022), "An image is worth one word: personalizing text-to-image generation using textual inversion", *arXiv preprint arxiv:2208.01618*.
- Gao, X., Gao, F., Tao, D. and Li, X. (2013), "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 24 No. 12.
- Ghadiyaram, D. and Bovik, A.C. (2015), "Massive online crowdsourced study of subjective and objective picture quality", *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, Vol. 25 No. 1, pp. 372-387.
- Ghildyal, A. and Liu, F. (2022), "Shift-tolerant perceptual similarity metric", *European Conference on Computer Vision, Springer*, pp. 91-107.
- Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y. and Cohen-Or, D. (2022), "Prompt-to-prompt image editing with cross attention control", *arXiv preprint arXiv:2208.01626*.
- Ho, J., Jain, A. and Abbeel, P. (2020), "Denoising diffusion probabilistic models", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840-6851.
- Hosu, V., Lin, H., Sziranyi, T. and Saupe, D. (2020), "KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment", *IEEE Transactions on Image Processing*, Vol. 29, pp. 4041-4056.
- Kang, L., Ye, P., Li, Y. and Doermann, D. (2014), "Convolutional neural networks for no-reference image quality assessment", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733-1740.
- Ke, J., Wang, Q., Wang, Y., Milanfar, P. and Yang, F. (2021), "MusiQ: multi-scale image quality transformer", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148-5157.

- Kumari, N., Zhang, B., Zhang, R., Shechtman, E. and Zhu, J.-Y. (2023), "Multi-concept customization of text-to-image diffusion", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931-1941.
- Li, J., Wang, Q., Qiu, J. and Dong, C. (2015), "Robust recovery of wideband block-sparse spectrum based on MAP and MMSE estimator", 2015 *IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings IEEE*, pp. 1783-1788.
- Li, C., Liu, L., Yu, B., Qiu, J. and Zhan, Y. (2025), "Re-initialization token learning for tool-augmented large language models", *arXiv preprint arXiv:2506.14248*.
- Ma, K., Liu, W., Liu, T., Wang, Z. and Tao, D. (. (2017), "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs", *IEEE Transactions on Image Processing*, Vol. 26 No. 8, pp. 3951-3964.
- Mittal, A., Moorthy, A.K. and Bovik, A.C. (2012a), "No-reference image quality assessment in the spatial domain", *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, Vol. 21 No. 12, pp. 4695-4708.
- Mittal, A., Soundararajan, R. and Bovik, A.C. (2012b), "Making a completely blind image quality analyzer", *IEEE Signal Processing Letters*, Vol. 20 No. 3, pp. 209-212.
- Mokady, R., Hertz, A., Aberman, K., Pritch, Y. and Cohen-Or, D. (2023), "Null-text inversion for editing real images using guided diffusion models", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038-6047.
- Moorthy, A.K. and Bovik, A.C. (2010), "A two-step framework for constructing blind image quality indices", *IEEE Signal Processing Letters*, Vol. 17 No. 5, pp. 513-516.
- Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al. (2015), "Image database TID2013: Peculiarities, results and perspectives", *Signal Processing: Image Communication*, Vol. 30, pp. 57-77.
- Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M. and Battisti, F. (2009), "TID2008-a database for evaluation of full-reference visual quality assessment metrics", *Advances of Modern Radioelectronics*, Vol. 10 No. 4, pp. 30-45.
- Qin, G., Hu, R., Liu, Y., Zheng, X., Liu, H., Li, X. and Zhang, Y. (2023), "Data-efficient image quality assessment with attention-panel decoder", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37 No. 2, pp. 2091-2100.
- Qiu, J. (2016), *Convolutional Neural Network Based Age Estimation From Facial Image and Depth Prediction From Single Image*, The Australian National University, Australia.
- Qiu, J., Dai, Y., Zhang, Y. and Alvarez, J.M. (2015), "Hierarchical aggregation based deep aging feature for age prediction", 2015 *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, pp. 1-5.
- Qiu, J., Wang, X., Fua, P. and Tao, D. (2019a), "Matching seqlets: an unsupervised approach for locality preserving sequence matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43 No. 2, pp. 745-752.
- Qiu, J., Wang, X., Maybank, S.J. and Tao, D. (2019b), "World from blur", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8493-8504.
- Qiu, J., Yang, Y., Wang, X. and Tao, D. (2020), "Hallucinating visual instances in total absentia", *European Conference on Computer Vision*, Springer, pp. 264-282.
- Qiu, J., Yang, Y., Wang, X. and Tao, D. (2021), "Scene essence", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8322-8333.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021), "Learning transferable visual models from natural language supervision", *International Conference on Machine Learning, PmlR*, pp. 8748-8763.

- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I. (2021), "Zero-shot text-to-image generation", *International Conference on Machine Learning, PmlR*, pp. 8821-8831.
- Ren, Y., Liu, L., Yu, B., Qiu, J. and Chen, Q. (2025a), "Instruction learning paradigms: a dual perspective on white-box and black-box LLMs", *arXiv preprint arXiv:2506.21573*.
- Ren, Y., Zhang, H., Wu, F., Qiu, J., Huang, J., Yu, B. and Liu, L. (2025b), "SIGMA: refining large language model reasoning via sibling-guided Monte Carlo augmentation", *arXiv preprint arXiv:2506.06470*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022), "High-resolution image synthesis with latent diffusion models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684-10695.
- Ruder, S. (2016), "An overview of gradient descent optimization algorithms", *arXiv preprint arXiv:1609.04747*.
- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M. and Aberman, K. (2023), "Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500-22510.
- Sheikh, H.R. and Bovik, A.C. (2006), "Image information and visual quality", *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, Vol. 15 No. 2, pp. 430-444.
- Sheikh, H.R., Sabir, M.F. and Bovik, A.C. (2006), "A statistical evaluation of recent full reference image quality assessment algorithms", *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, Vol. 15 No. 11, pp. 3440-3451.
- Song, J., Meng, C. and Ermon, S. (2020), "Denoising diffusion implicit models", *arXiv preprint arXiv:2010.02502*.
- Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J. and Zhang, Y. (2020), "Blindly assess image quality in the wild guided by a self-adaptive hyper network", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667-3676.
- Sun, W., Min, X., Tu, D., Ma, S. and Zhai, G. (2023), "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 17 No. 6, pp. 1178-1192.
- Sun, W., Xie, J., Qiu, J. and Ma, Z. (2021), "Part uncertainty estimation convolutional neural network for person re-identification", *2021 IEEE International Conference on Image Processing (ICIP), IEEE*, pp. 2304-2308.
- Tang, S., Xu, X., Qiu, J. and Wang, Z. (2026), "Controllable data generation with hierarchical neural representations", in *Forty-second International Conference on Machine Learning*.
- Van Den Oord, A. and Vinyals, O., *et al* (2017), "Neural discrete representation learning", *Advances in Neural Information Processing Systems*, Vol. 30.
- Wang, Z., Simoncelli, E.P. and Bovik, A.C. (2003), "Multiscale structural similarity for image quality assessment", in *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, IEEE*, Vol. 2, pp. 1398-1402.
- Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004), "Image quality assessment: from error visibility to structural similarity", *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, Vol. 13 No. 4, pp. 600-612.
- Wang, Y., Xu, C., Qiu, J., Xu, C. and Tao, D. (2018), "Towards evolutionary compression", *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2476-2485.
- Wang, H., Liu, J., Tan, H., Lou, J., Liu, X., Zhou, W. and Liu, H. (2024), "Blind image quality assessment via adaptive graph attention", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 34 No. 10, pp. 10299-10309.

- Xu, X., Qiu, J., Wang, X. and Wang, Z. (2022), "Relationship spatialization for depth estimation", European Conference on Computer Vision, *Springer*, pp. 615-637.
- Xu, X., Qiu, J., Yu, B. and Wang, Z. (2024), "Visual relationship transformation", European Conference on Computer Vision, *Springer*, pp. 251-272.
- Yang, Y., Qiu, J., Song, M., Tao, D. and Wang, X. (2020a), "Distilling knowledge from graph convolutional networks", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7074-7083.
- Yang, Y., Qiu, J., Song, M., Tao, D. and Wang, X. (2020b), "Learning propagation rules for attribution map generation", European Conference on Computer Vision, *Springer*, pp. 672-688.
- Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J. and Yang, Y. (2022), "Maniqa: multi-dimension attention network for no-reference image quality assessment", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191-1200.
- Ye, P. and Doermann, D. (2012), "No-reference image quality assessment using visual codebooks", *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, Vol. 21 No. 7, pp. 3129-3138.
- Ye, P., Kumar, J. and Doermann, D. (2014), "Beyond human opinion scores: blind image quality assessment based on synthetic scores", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4241-4248.
- Ye, P., Kumar, J., Kang, L. and Doermann, D. (2012), "Unsupervised feature learning framework for no-reference image quality assessment", 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, *IEEE*, pp. 1098-1105.
- Yin, G., Wang, W., Yuan, Z., Han, C., Ji, W., Sun, S. and Wang, C. (2022), "Content-variant reference image quality assessment via knowledge distillation", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36 No. 3, pp. 3134-3142.
- Yu, H., Li, R., Xie, S. and Qiu, J. (2024), "Shadow-enlightened image outpainting", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7850-7860.
- Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O. (2018), "The unreasonable effectiveness of deep features as a perceptual metric", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586-595.
- Zhang, W., Ma, K., Yan, J., Deng, D. and Wang, Z. (2020), "Blind image quality assessment using a deep bilinear convolutional neural network", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 30 No. 1, pp. 36-47, doi: [10.1109/TCSVT.2018.2886771](https://doi.org/10.1109/TCSVT.2018.2886771).
- Zhang, H., Liu, L., Yu, B., Qiu, J., Ren, Y. and Liu, X. (2025a), "LARGO: low-rank regulated gradient projection for robust parameter efficient fine-tuning", *arXiv preprint arXiv:2506.12394*.
- Zhang, H., Liu, L., Yu, B., Qiu, J., Xiao, L., Ren, Y., Chen, Q. and Liu, X. (2025b), "ContextPRM: leveraging contextual coherence for multidomain test-time scaling", *arXiv preprint arXiv:2509.24460*.
- Zhu, H., Li, L., Wu, J., Dong, W. and Shi, G. (2020), "MetaIQA: deep meta-learning for no-reference image quality assessment", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14143-14152.

Corresponding author

Shan Huang can be contacted at: shan.huang@blcu.edu.cn