Structured Creative Evaluation for Text-to-Image Generative AI Models

Wancheng Ma[1], Qiong Zhang[2]
[1]School of Information Engineering, Zhijiang College of Zhejiang University of Technology, Shaoxing 312030, China
[2]School of Computer and Information Engineering, Bengbu University, Bengbu 233030, China
Corresponding author: Qiong Zhang (e-mail: qiong.zhang.1@outlook.com).

ABSTRACT In recent years, text-to-image (T2I) generation models have made substantial progress, particularly in visual realism and the expression of prompt semantics. However, a key difficulty remains: how to evaluate generated results automatically in a way that is both comprehensive and interpretable, while still being practical for real deployment. To address this issue, this paper proposes a multi-dimensional image quality assessment framework for T2I tasks. The framework examines generated images from five dimensions—text fidelity, perceptual quality, object consistency, relational consistency, and global semantic alignment—and derives a final quality score through normalization and weighted fusion. In terms of methodology, the framework combines Tesseract OCR, perceptual quality analysis based on Laplacian variance and exposure statistics, YOLO object detection, BLIP-based visual question answering, and CLIP image-text similarity, thereby forming a modular evaluation pipeline with diagnostic capability. Experiments on multiple mainstream T2I models and representative prompts show that the proposed method can not only distinguish overall performance differences across models, but also provide interpretable results at the level of individual dimensions.

INDEX TERMS Structured creative evaluation, text-to-image generation, image quality assessment, visual question answering, object consistency

I. INTRODUCTION

With the rapidly increasing development of generative models, in particular, text-to-image (T2I) generation has become a key task in the field of multimodal AI, which directly helps bridge natural language understanding and visual content synthesis. The development of early generative adversarial networks (GANs) provided a crucial foundation for high-quality image synthesis by introducing the adversarial training methodology, which greatly improved visual fidelity and diversity in the generated samples [1]. To further enhance the scalability and controllability of the generation process, latent diffusion models (LDMs) operated in a compressed latent space and made it both feasible and computationally less expensive to generate high-resolution images [2]. This brought about widespread applications of text-to-image (T2I) systems in research and development. Recently, large-scale models like DALL·E 2 and Imagen have extended these advances with excellent semantic understanding, compositional reasoning, and photorealistic image generation, thus establishing new baselines for text-to-image (T2I) systems [3,4]. Despite these substantial improvements in both generation quality and controllability, the challenge of T2I model evaluation persists as a fundamentally complex issue, not yet fully resolved. The conventional metrics for image quality assessment, like structural similarity (SSIM) and peak signal-to-noise ratio (PSNR), are mainly designed for distortion-based evaluation scenarios, where a reference image is available. Consequently, they are not well-suited for generative tasks, particularly in T2I settings where the primary goal is not to reconstruct a ground-truth image but rather to generate a semantically consistent and visually plausible image conditioned on text descriptions [5]. These metrics fail to capture critical aspects such as semantic alignment, object correctness, and relational consistency between the textual input and generated visual content. To overcome this limitation and address such scenarios, generative model evaluation metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) have gained popularity. IS assesses the sample's diversity and confidence through classification entropy, while FID measures the distributional distance between generated images and real images in feature space. Although these metrics provide useful insights into the overall realism and diversity of generated datasets, they operate primarily at a global distribution level and lack the granularity to assess fine-grained semantic correctness. Specifically, they are not useful for identifying errors in object attributes, spatial relationships, or textual inconsistencies within an image, which are crucial to evaluating T2I systems [6,7]. With the advent of large-scale vision-language models (VLMs) in recent years, there has been a rise in exploring automatic and semantically aware evaluation for T2I systems. VLMs like CLIP can measure

global semantic alignment by projecting both images and text into a common embedding space and comparing the similarities that represent high-level correspondence [8]. On the other hand, visual understanding and multimodal reasoning models like BLIP and BLIP-2 show significant prowess that can be exploited for further evaluation [9,10]. Inspired by these capabilities, we introduce TIFA, which follows a principled evaluation methodology by automatically generating question-answer pairs from a given textual prompt using BLIP, and employing BLIP and BLIP-2 as a VQA system to determine whether the generated image reflects the content described in the text [11]. Additionally, some recent benchmarks and evaluation frameworks have emerged to enhance the T2I evaluation landscape. T2I-CompBench focuses on compositional reasoning and tests the model's ability to accurately represent intricate attribute-object and object-object relationships. GenEval emphasizes object-level alignment and examines the presence and correctness of specific entities mentioned in the text. ImageReward leverages a learned reward model based on human preferences to better reflect subjective evaluation. VQAScore also quantifies the semantic correctness of generated images through a question-answering approach, and STRICT particularly aims to evaluate the quality of text rendering in images, thus addressing a significant weakness in many generative models [12-15]. All these research directions shed light on the growing trend toward more fine-grained, interpretable, and task-specific evaluation protocols. However, current evaluation methods still present limitations in terms of coverage of semantic dimensions and reliability due to proxy models having their own biases, as well as a lack of interpretability in diagnosing failure modes. Hence, designing a comprehensive, reliable, and scalable evaluation framework for text-to-image generation systems is still an open and crucial research task. This challenge is a compelling reason to further investigate structured, multi-dimensional, and human-aligned approaches to evaluation methodology for text-to-image generation systems. Although recent advances in neural image generators have achieved impressive generation quality, the evaluation methods used to quantify generation quality still fall short of many requirements and practices in real-world use cases. This paper identifies the following issues in current evaluation methods: low interpretability, failure to jointly model semantic constraints, and low diagnostic capability for single images. In light of the limitations of current approaches, this paper proposes a novel multi-dimensional, interpretable, and practical framework for text-to-image (T2I) IQA. The key insight of the proposed framework is to decompose the evaluation procedure into multiple explicitly defined dimensions so that various aspects of model generation quality can be evaluated in an organized way instead of being reduced to a single scalar value. With this novel IQA framework, models can be evaluated more transparently and systematically. First, the whole framework is developed using multiple existing modules, which can be simply integrated in an orderly manner to fit a given pipeline. Unlike introducing some sophisticated yet huge models, the proposed approach focuses on ways to combine these proven elements in a coherent and understandable manner. Thus, each evaluation dimension can be mapped to an individual functional component, so the whole system can be easily analyzed and reproduced. The main contributions of this paper are as follows:

(1). A five-dimensional evaluation system is built, including OCR text fidelity, perceptual quality, object consistency, relation consistency, and global semantic alignment. Different aspects of T2I generation quality are covered, allowing for a more fine-grained assessment than a single metric.

(2). We build an automatic evaluation pipeline by integrating Tesseract, YOLO, BLIP, and CLIP into a unified framework, so that each module is responsible for one evaluation dimension as part of an interpretable evaluation procedure with a clear decomposition.

(3). Based on experimental results on multiple models, as well as parameter sensitivity analysis and representative visual cases, the effectiveness of the proposed method in distinguishing generation quality and providing analysis from different dimensions is demonstrated.

II. RELATED WORK

Recent years have seen continuous progress in research on text-to-image (T2I) generation and its automatic evaluation. The state-of-the-art research can be generally summarized in three main directions: First, generative models have evolved from the GAN-based paradigm to diffusion-based paradigms, resulting in continuous improvement in terms of image realism, resolution, and controllability under textual conditions. Second, traditional methods for image quality assessment (IQA) have gradually evolved from full-reference settings to no-reference settings, with increasing focus on perceptual and natural scene statistics fidelity, as well as visual aesthetic quality. Third, and most importantly, emerging automatic evaluation approaches for T2I tasks have gradually turned toward vision-language models, compositional semantic constraints, and preference-learning mechanisms, and have shifted the evaluation objective from merely assessing visual realism to evaluating the faithful satisfaction of given textual instructions. However, as a result of this focus, most existing methods only prioritize one aspect or a small subset of aspects (e.g., perceptual quality, global semantic alignment, object-level correctness, text rendering performance, etc.). This limits their capability to provide a comprehensive and fine-grained evaluation of generated images. Notably, none of the current approaches has created a unified framework that can simultaneously account for local visual details, relational constraints, explicit text content, and overall semantic consistency, while being sufficiently interpretable and practical for real-world applications [5-18]. In this work, we propose a new approach for text-image alignment that combines visual perceptual quality, semantic consistency, and both text-to-image and image-to-text generation.

A. TEXT-TO-IMAGE GENERATION MODELS

T2I models like Generative Adversarial Networks (GANs) have come a long way from the days when they produced high-quality images using adversarial training involving a generator and a discriminator [1]. Nevertheless, the training for these networks is often unstable and cannot be scaled up to generate high-resolution images. Unlike GANs, Latent Diffusion Models (LDMs) perform diffusion in latent spaces, allowing cheaper computation for generating high-resolution images [2]. These advancements allow for the development of new models such as DALL·E 2 and Imagen that can provide enhanced visual quality and understanding of semantically complex prompts. This shows that T2I models are becoming increasingly capable not only of generating photorealistic images but also of comprehending textual prompts [3, 4].

B. IMAGE QUALITY ASSESSMENT AND AUTOMATIC T2I EVALUATION

Traditional image quality assessment methods, including both full-reference and no-reference approaches such as SSIM, PSNR, BRISQUE, NIQE, NIMA, and MUSIQ, mainly evaluate perceptual fidelity, natural scene statistics, and visual aesthetic quality [5, 16, 17, 19]. These methods are effective for general-purpose image quality evaluation, but they were not designed for text-conditioned image generation. To narrow this gap, CLIP-IQA introduces vision-language priors into no-reference IQA, offering a useful direction for linking perceptual quality assessment with semantic understanding. For T2I evaluation, metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) are still widely used in automatic evaluation [6, 7]. However, both mainly operate at the distribution level and therefore cannot directly judge whether a single generated image satisfies the object-level, attribute-level, or relational constraints specified in the input text. In recent years, the development of methods based on vision-language models and compositional semantics constraints has gained significant attention. Some well-known methods can be found, such as CLIPScore, TIFA, T2I-CompBench, GenEval, ImageReward, VQAScore, and STRICT [11-15, 18]. Such methods can improve image-text alignment from different perspectives. Although these methods have their advantages, a single metric can only capture part of the defects in generated images, such as object identification, attribute consistency, and the accuracy of relations. In the case of an image containing explicit text, the newly developed STRICT method provides a more comprehensive evaluation of the rendered text from various perspectives, such as readability, correctness, and adherence to the instructions given in the prompt.

C. LIMITATIONS OF EXISTING METHODS

Despite these advances, most existing methods still concentrate on only one aspect, or at best a small subset of aspects, such as perceptual quality, global semantic alignment, object-level correctness, or text rendering performance. As a result, they often cannot provide a comprehensive and fine-grained evaluation of generated images. More specifically, current approaches still have difficulty capturing, within a single framework, the different types of constraints implied by text prompts, including explicit text correctness, object presence and attributes, inter-object relationships, perceptual visual quality, and overall image-text semantic consistency. In addition, many methods depend on single summary scores or black-box models, which weakens their interpretability and limits their usefulness for diagnosis. These limitations point to the need for a unified and interpretable evaluation framework that can separate different evaluation dimensions while still providing both an overall assessment and fine-grained diagnostic insight.

III. METHOD

To address the above limitations, particularly the lack of unified and interpretable multi-dimensional evaluation, this paper proposes a multi-dimensional image quality assessment method for T2I-generated results, with two primary goals: unified modeling and interpretable diagnosis. Unlike evaluation approaches that rely on a single metric or a single analytical pathway, the proposed method explicitly decomposes different types of prompt constraints into computable evaluation targets and analyzes generated images from five perspectives: text correctness, perceptual visual quality, object presence and category consistency, relational satisfaction among objects, and global image-text semantic alignment. From a methodology standpoint, the framework explicitly uses modules that have tangible or definable meaning to ensure that the evaluation is transparent, the results from each dimension are understandable, and the whole process is replicable and scalable. In terms of output, the method not only produces a unified final score but also preserves diagnostic results for each dimension, making it better suited to model comparison, failure-case analysis, and sampling-parameter tuning in both research and practical deployment settings [8-14, 18, 20].

A. OVERALL FRAMEWORK

The core idea of the proposed method is as follows: first, explicit text, target objects, relational question-answer pairs, and semantic targets required for evaluation are constructed according to the prompt; then, multimodal parsing is performed on the generated image; finally, five dimension-wise sub-scores are obtained and normalized and fused with weights to output the final quality score and dimension-wise results. Fig. 1 shows the overall structure of the proposed assessment framework.

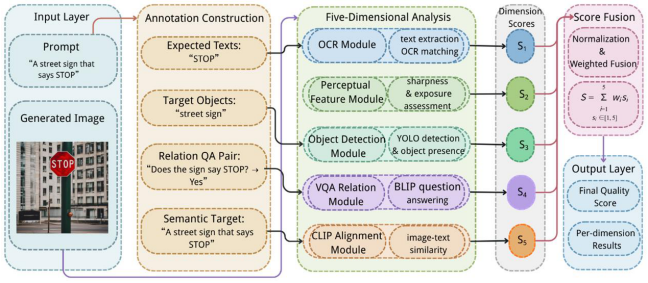

FIGURE 1. Multi-dimensional T2I image quality assessment framework proposed in this paper. The framework constructs explicit text, target objects, relational question-answer pairs, and semantic targets from the prompt, and obtains scores for five dimensions through OCR, perceptual quality analysis, object detection, VQA, and CLIP modules, respectively. Finally, the overall score and dimension-wise results are produced after normalization and weighted fusion.

As shown in Fig. 1, the architecture consists of an Input Layer, an Annotation Construction layer, a Five-Dimensional Analysis layer, a Score Fusion layer, and an Output Layer. Among them, the Annotation Construction layer shows an important distinction between the proposed method and traditional single-metric evaluation, because it explicitly transforms the prompts into computable evaluation targets, enabling the OCR, object detection, and visual question answering modules to analyze the image around the same prompt semantics.

B. DEFINITIONS AND SCORING CRITERIA OF THE FIVE ATTRIBUTES

To enhance the interpretability and reproducibility of the five-dimensional evaluation system, this paper first clarifies the semantic connotations of the five attributes and then provides unified scoring criteria on a 1-5 scale. This design makes it convenient to incorporate local text, visual quality, object existence, relational constraints, and overall semantic alignment into the same evaluation framework, and also provides a clear semantic basis for subsequent mathematical modeling.

Table 1 summarizes the functional roles of the five attributes in this paper. Among them, text fidelity mainly corresponds to OCR-driven text rendering evaluation; object consistency and relational consistency respectively draw on ideas from object detection, compositional semantic evaluation, and VQA; and global semantic alignment borrows from CLIP-style image-text alignment methods. Therefore, Table 1 is not only the semantic definition table of the evaluation dimensions, but also the conceptual basis for subsequent formula modeling [8-13, 18, 20].

Table 2 maps the five attributes to a unified 1-5 point evaluation range, where 1 point indicates severe failure, 3 point indicates basic usability, and 5 point indicates high consistency with the prompt. This unified scale allows the outputs of different modules to be directly weighted during the fusion stage and also makes comparisons across dimensions and case diagnosis more intuitive. In terms of evaluation philosophy, this table is consistent with existing works on perceptual quality, image-text alignment, object-level evaluation, and question-answering-based faithfulness evaluation, while placing greater emphasis on unified scoring and an interpretability loop in engineering practice [5, 8, 11- 13, 18].

C. MATHEMATICAL MODELING AND SCORING FUNCTIONS OF THE FIVE ATTRIBUTES

Based on the definitions and scoring criteria above, the five attributes are further formalized as computable scoring functions.

1) Text Fidelity

Let the expected text set be $E = \{e_i\}_{i=1}^{m}$, and let the candidate text extracted from the image by OCR be $R = \{r_j\}_{j=1}^{m}$. First, the text is normalized (lowercasing, punctuation removal, and whitespace normalization). For any text pair, the character-level similarity and token-level similarity are defined as follows

$$\text{char\_sim}(a, b) = \frac{2M(a, b)}{|a| + |b|} \tag{1}$$

$$\text{token\_sim}(a, b) = \frac{|\omega(a) \cap \omega(b)|}{|\omega(a) \cup \omega(b)|} \tag{2}$$

This design is similar to the idea of Tesseract OCR. Here, $M(a, b)$ denotes the number of matched characters obtained by the sequence matching algorithm, $|a|$ and $|b|$ denote the lengths of strings a and b, respectively, and $\omega(\cdot)$ denotes the token set after word segmentation. The overall similarity is defined as

$$\text{sim}(a, b) = 0.6\,\text{char\_sim}(a, b) + 0.4\,\text{token\_sim}(a, b) \tag{3}$$

If OCR results contain confidence scores $c_j \in [0, 1]$, then, for each expected text item e_i, the optimal matching score is defined as

$$P_i = \max_{r_j \in R} (\text{sim}(e_i, r_j) \cdot (0.6 + 0.4c_j)) \tag{4}$$

The text fidelity score can then be written as:

$$S_1 = 1 + 4 \cdot \frac{1}{m} \sum_{i=1}^{m} P_i \tag{5}$$

When a sample does not contain explicit text annotations, a full score of 5.0 is returned by default in engineering practice to avoid incorrectly penalizing irrelevant samples. This treatment is intended to maintain a consistent scale across the full dataset, but when reporting results on the text subset, explicit-text samples should still be discussed separately.

TABLE 1. Meanings of the five attributes.

Attribute	Core Meaning	Evaluation Role
Text Fidelity	Evaluates the consistency between explicit text in the image and the target text at the character level, token level, and readability level.	Measures the text rendering capability of T2I models, especially for text-in-image scenarios.
Perceptual Quality	Evaluates image sharpness, exposure reasonableness, and overall visual readability.	Reflects the basic visual quality and generation stability of generated images.
Object Consistency	Evaluates whether key objects in the prompt truly appear and whether their categories and counts match.	Distinguishes between "looks like" and "is correct," emphasizing object presence.
Relational Consistency	Evaluates whether spatial, attribute, and semantic relations among objects satisfy prompt constraints.	Measures the model's ability to understand relations, which is a major challenge in complex T2I tasks.
Global Semantic Alignment	Evaluates the match between the whole image and the prompt text at the overall semantic level.	Compensates for the limitations of local metrics and avoids cases where "local parts are correct but the overall image misses the point."

TABLE 2. Unified scoring criteria of the five attributes (1–5).

Attribute	1 point	2 points	3 points	4 points	5 points
Text Fidelity	Garbled or completely unreadable	Severe spelling errors or truncation	Multiple spelling errors but still recognizable	A few minor spelling errors that do not affect understanding	All text is completely correct and clearly readable
Perceptual Quality	Severely distorted and difficult to recognize	Obvious blur or compression artifacts	Moderately blurry but recognizable	Slight blur or minor noise	Clear and sharp, with no obvious noise
Object Consistency	Many missing or incorrect objects	Multiple object errors	A few key object errors	A count or category deviation in one object	All key object categories and counts are completely correct
Relational Consistency	Relations are completely wrong	Most relations are wrong	Some relations are wrong	One minor relational deviation	All spatial/semantic relations are correct
Global Semantic Alignment	Clearly irrelevant to the prompt	Only partially relevant	Generally relevant but missing key elements	Basically consistent, with slight detail deviations	The overall content, style, and semantics of the image all match the prompt

2) Perceptual Quality

Perceptual quality is used to measure the sharpness and exposure reasonableness of generated images at the visual level. In this paper, this dimension is composed of two weighted sub-items: sharpness and exposure. Let the grayscale image be G. Compared with methods such as $v_{\text{lap}} = \text{Var}(\nabla^2 G)$, NIMA, MUSIQ, BRISQUE, NIQE and CLIP-IQA, the sharpness intensity is defined based on Laplacian variance here emphasizing a lightweight, interpretable, and training-free engineering implementation [16, 17]. To map it to the unified five-point scale, it is defined as

$$q_{\text{sharp}} = 1 + 4 \cdot \text{clip}\left(\frac{v_{\text{lap}} - 50}{950}, 0, 1\right) \tag{6}$$

Here, the thresholds "50" and "950" are normalization boundaries empirically set according to the distribution of experimental samples. Furthermore, let the average image brightness be μ, since the 8-bit grayscale range is [0, 255], the midpoint 127.5 can be regarded as the ideal exposure center; therefore, exposure reasonableness is defined as:

$$q_{\text{exp}} = 1 + 4 \cdot \text{clip}\left(1 - \frac{|\mu - 127.5|}{127.5}, 0, 1\right) \tag{7}$$

Here, the clipping function is defined as

$$\text{clip}(x, 0, 1) = \min(1, \max(0, x)) \tag{8}$$

The final perceptual quality score is

$$S_2 = 0.7 q_{\text{sharp}} + 0.3 q_{\text{exp}} \tag{9}$$

While keeping the computation simple, this design has good interpretability and is suitable as the basic visual quality component in the multi-dimensional framework [16, 17].

3) Object Consistency

Object consistency is used to assess whether target objects exist in the generated image and whether they are consistent with the expected description at the category level. Let the expected object set be $O = \{o_i\}_{i=1}^{n}$, and let the output of the object detector on image I be denoted by $D(I) = \{c_j, p_j\}_{j=1}^{N}$, where c_j and p_j denote the detected class and its confidence score, respectively. For any category c, the maximum detection confidence is defined as

$$\text{Conf}(c) = \max_{j: c_j = c} p_j \tag{10}$$

This dimension is consistent with the evaluation goals emphasized by T2I-CompBench and GenEval regarding object presence, counting, and attribute consistency [12, 13]. Considering that object names in prompts may not completely match detector category labels, this paper first uses a mapping function $m(o)$ to map target objects to common detector category names. If $m(o)$ can be directly matched in the detection results, the object score is defined as: $\text{Strength}_o = \text{Conf}(m(o))$. If no direct match exists, the name similarity between detected categories and $m(o)$ is further computed. When $\text{Sim}(c, m(o)) \geq \tau$, an approximate matching score is adopted as

$$\text{Strength}_o = \text{Conf}(c) \cdot \text{Sim}(c, m(o)) \tag{11}$$

The threshold is used to control the strictness of approximate name matching and is empirically set to 0.72 based on experimental observations. The detector component adopts a YOLO-style object recognition idea to improve the computability of object presence [20].

If no match can still be found, the VQA model is used to answer the question "Is there a o?", and the result is mapped to object strength: answer "yes" takes 0.6, answer "no" takes 0, and other cases take 0.2. Finally, the overall object consistency strength is defined as

$$\text{Strength} = \frac{1}{n}\sum_{i=1}^{n}\text{Strength}_{o_i} \tag{12}$$

and is mapped to a five-point score using

$$S_3 = 1 + 4 \cdot \text{Strength} \tag{13}$$

4) Relational Consistency

Relational consistency is used to evaluate whether the generated image satisfies the relational constraints in the prompt or annotations. For each relational question, let the input image be I and the question be q. The VQA model outputs the predicted answer a, and a^* denotes the corresponding expected answer. To reduce the influence of differences in case, punctuation, and whitespace, the answers are first uniformly normalized. For each answer pair, the character-level similarity and token-level similarity are defined as follows

$$\text{sim}_{\text{char}}(a, a^*) \in [0, 1] \tag{14}$$

$$\text{sim}_{\text{token}}(a, a^*) = \frac{|\omega(a) \cap \omega(a^*)|}{|\omega(a) \cup \omega(a^*)|} \tag{15}$$

This dimension shares the same evaluation goal as methods such as TIFA, T2I-CompBench, and VQAScore in emphasizing relations, attributes, and fine-grained instruction-following ability [11, 12]. Here, $\omega(\cdot)$ denotes the normalized token set. When the denominator is zero, let $\text{sim}_{\text{token}}(a, a^*) = 0$; the overall similarity can be defined as

$$\text{sim}(a, a^*) = 0.6\,\text{sim}_{\text{char}}(a, a^*) + 0.4\,\text{sim}_{\text{token}}(a, a^*) \tag{16}$$

For general open-ended answers, the score is defined as $\text{Score} = 1 + 4 \cdot \text{sim}(a, a^*)$. For yes/no questions, if the predicted answer is exactly the same as the expected answer, Score = 5; if their polarities are opposite, Score = 1; otherwise, the score is still computed according to text similarity. Finally, the relational consistency score is defined as

$$S_4 = \min(5, \max(1, \text{Score})) \tag{17}$$

5) Global Semantic Alignment

Global semantic alignment is used to evaluate the overall semantic match between the generated image and the input text. Let the input image be I, and the input text be T. A pre-trained CLIP model is used to extract image features and text features, and we have $f_I = E_I(I)$ and $f_T = E_T(T)$, where $E_I(\cdot)$ and $E_T(\cdot)$ denote the CLIP image encoder and text encoder, respectively. This dimension is close in objective to overall semantic evaluation ideas such as CLIPScore and VQAScore, but in this paper it is uniformly incorporated into the interpretable five-dimensional framework [8, 18]. The semantic similarity between the image and the text is computed using cosine similarity as

$$S_{\text{clip}} = \frac{f_I \cdot f_T}{\|f_I\| \cdot \|f_T\|} \tag{18}$$

To keep it consistent with the other evaluation dimensions, $S_{\text{clip}} \in [-1, 1]$ is first normalized to the [0,1] interval using $S_{\text{norm}} = (S_{\text{clip}} + 1)/2$. Then, the normalized result is mapped to the five-point scoring interval to obtain the global semantic alignment score by $S_{\text{five}} = 1 + 4S_{\text{norm}}$. Finally, to ensure that the score lies within the range, it is defined as

$$S_5 = \min\left(5, \max\left(1, 1 + 4 \cdot \frac{S_{\text{clip}} + 1}{2}\right)\right) \tag{19}$$

6) Final Score

Let S_1, S_2, S_3, S_4, and S_5 denote the scores of the five attributes, respectively. The final score is defined as

$$S = 0.15S_1 + 0.2S_2 + 0.2S_3 + 0.2S_4 + 0.25S_5 \tag{20}$$

where the coefficients are empirically determined based on practical considerations. The proposed five-dimensional framework combines completeness and interpretability: CLIP provides global semantic constraints, OCR handles explicit text, YOLO and VQA cover objects and relations respectively, and perceptual quality compensates for the lack of attention paid by purely semantic evaluation to the visual quality of images. This design forms a complementary relationship with existing compositional semantic evaluation, text rendering evaluation, and preference-based overall evaluation methods [11-15, 18].

IV. EXPERIMENTS AND RESULTS ANALYSIS

To systematically validate the effectiveness of the proposed five-dimensional evaluation framework, this paper conducts experimental analysis from four aspects: model discriminability, parameter sensitivity, sample-level diagnostic capability, and consistency with existing automatic evaluation methods. Specifically, the experiments not only examine the differences in overall scores among multiple T2I models under a unified evaluation protocol, but also further analyze how dimension-wise scores vary for the same model under different sampling configurations, together with visual cases to investigate the correspondence between quantitative scores and actual generation quality. In addition, several representative automatic evaluation methods are selected as nearby baselines for comparison from the perspectives of text fidelity, semantic alignment, and overall evaluation, so as to verify the rationality of the proposed framework

in trend consistency, error localization, and comprehensive analysis. Through this multi-level experimental design, our work demonstrates that the proposed method can not only distinguish overall performance differences across models, but also provide finer-grained and more interpretable evidence for identifying the sources of failure [11-15, 18].

A. EXPERIMENTAL SETTINGS

This paper selects six mainstream T2I models for experiments: SD-XL, SD-v1-4, SD2.1, CogView4, BLIP3o- 4B, and OpenJourney. All results are computed under the unified five-dimensional evaluation framework, and the final score is the average of the five sub-scores. Table 3 presents the latest quantitative results across models.

As can be seen from Table 3, SD-XL ranks first with a final score of 3.7407, indicating that it achieves a good balance among perceptual quality, object consistency, and relational consistency. CogView4 performs strongly in text faithfulness, but its final score is slightly lower than that of SD-XL due to perceptual quality. BLIP3o-4B has better perceptual quality than most non-SD-XL models, but its object consistency is relatively low. OpenJourney obtains the lowest final score, mainly because of relatively weak perceptual quality and explicit text performance.

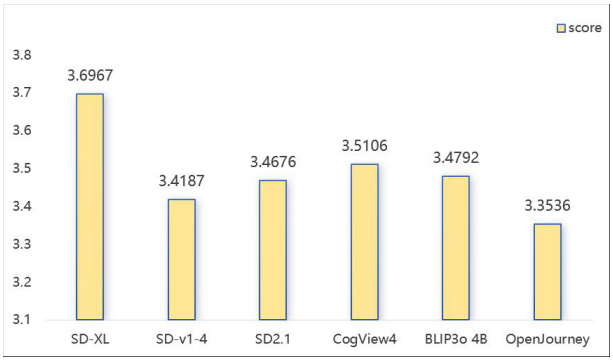

FIGURE 2. Bar chart of final scores for different T2I models.

B. SD-XL PARAMETER SENSITIVITY ANALYSIS

To further analyze the influence of different sampling configurations on the scoring results of the same model, this paper takes SD-XL as an example and compares the five-dimensional scores and final scores under different settings, including random seeds, numbers of generated samples, guidance scale (GS), and inference steps. Table 4 presents the scores under different parameter settings, and Fig. 3 shows the changing trends of perceptual quality and final score under different steps.

As can be seen from Table 4, the overall score of SD-XL under different settings shows a certain pattern. First, the final scores of w/diff seeds and Baseline are 3.7388 and 3.7407, respectively, with only a very small difference, indicating that under the current task setting, the influence of random seeds on the final score is relatively limited, although slight fluctuations can still be observed in object consistency and relational consistency. Second, w/25 images improve Text Fidelity and Relational Consistency, indicating that increasing the number of samples under the same prompt improves the probability of obtaining high-scoring samples, but its Object Consistency decreases slightly, suggesting that more samples do not necessarily bring stable improvements at the object level. Thirdly, the guidance scale is the parameter that shows the highest sensitivity. Specifically, at GS = 12, the final score obtains a maximum value equal to 3.7871. In the case where GS = 1, the final score decreases to 3.5339, while object and relational consistencies decline dramatically. The results demonstrate that textual guidance plays an essential role in meeting the constraints imposed by semantics. Weak guidance leads to a lack of constraint satisfaction.

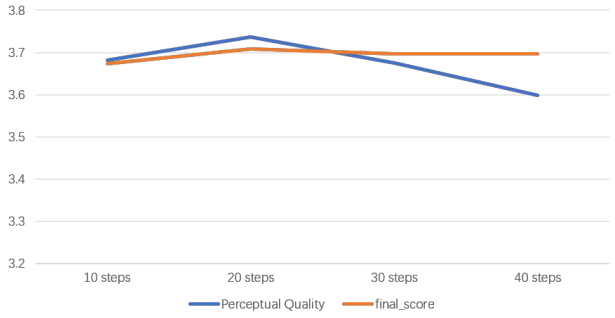

FIGURE 3. Trends of perceptual quality and final score under different inference steps.

As seen in Fig.3, an increase in the number of steps from 10 to 20 leads to a significant improvement in perceptual quality, which implies that an increase in denoising steps could help recover fine detail information and improve image sharpness. Nevertheless, further increases in the number of steps to 30 and 40 steps lead to deterioration in perceptual quality, implying that further increases in steps might not necessarily lead to better visual outcomes. On the contrary, in contrast to perceptual quality, the final score fluctuates slightly with varying numbers of steps, maintaining a high value at approximately 20 steps and remaining in a similar range. The reason for this phenomenon is probably that the final score is not solely dependent on perceptual quality but also relies on other metrics such as object consistency, relational consistency, and global semantic consistency.

C. COMPARISON OF VISUAL RESULTS AND SCORES UNDER DIFFERENT PROMPTS

In order to better illustrate the capability of diagnosis of our approach, Fig. 4 is an example illustrating the generated outputs and corresponding scores of several models under various prompts. Our proposed approach differs from the previous automatic metrics that are based solely on one metric, in that it relates visual output with numerical score, thereby enabling one to diagnose strengths and weaknesses of models under certain prompts.

TABLE 3. Scoring results of different T2I models across five dimensions.

Model	Text Fidelity	Perceptual Quality	Object Consistency	Relational Consistency	Global Semantic Alignment	Final Score
SD-XL	3.5808	3.6752	3.9252	3.9297	3.5903	3.7407
SD-v1-4	3.5918	2.8175	3.8086	3.5305	3.5627	3.4607
SD2.1	3.5449	2.9458	3.8074	3.6982	3.5593	3.5118
CogView4	3.6564	2.7897	3.8910	3.8454	3.5881	3.5507
BLIP3o-4B	3.4579	3.3395	3.6503	3.5877	3.5782	3.5287
OpenJourney	3.5284	2.3844	3.8289	3.6655	3.5784	3.3996

TABLE 4. Five-dimensional scores and final scores of SD-XL under different parameter settings.

Model	Text Fidelity	Perceptual Quality	Object Consistency	Relational Consistency	Global Semantic Alignment	Final Score
Baseline	3.5808	3.6752	3.9252	3.9297	3.5903	3.7407
w/diff seeds	3.5822	3.6353	3.9629	3.9208	3.5905	3.7388
w/25 images	3.7059	3.6135	3.8228	3.9649	3.5776	3.7305
w/gs 12	3.6041	3.7862	3.9483	4.0075	3.5926	3.7871
w/gs 5	3.5642	3.5339	3.9507	3.9018	3.5879	3.7089
w/gs 1	3.4863	3.6716	3.4048	3.5781	3.5202	3.5339

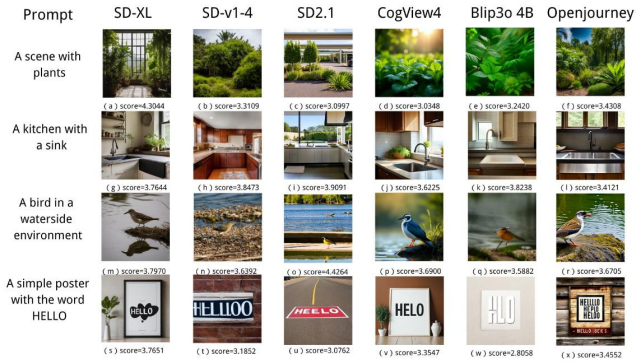

FIGURE 4. Generated results and corresponding final scores of different models under different prompts. Each row corresponds to one prompt, and each column corresponds to one generation model.

As can be observed from Figure 4, the findings obtained at the sample level also demonstrate the efficacy of our framework. The table shows the images generated along with the final scores obtained for each model under various prompt classes. On the whole, there exist significant distinctions among the models concerning tasks related to open scene, object constraint, environment relation, and explicit text generation. For those prompts which are scene based, like plant scene, kitchen scene etc., majority of models are capable of producing images which are semantically sensible but there is still variation among them in terms of subject salience, scene consistency, and naturalness and hence varied scores. For prompts containing both subject and environment relational constraints, the differences among models in semantic consistency between objects and environments are more pronounced, mainly reflected in dimensions such as object consistency, relational consistency, and global semantic alignment. For prompts containing explicit text content, the capability gap among models is the most obvious: some models can generate relatively clear and recognizable target text, while others suffer from spelling errors, repeated characters, or missing glyph components, leading to a significant decline in Text Fidelity and further affecting the final score. Overall, this group of visualization results shows that the proposed multi-dimensional evaluation framework can not only reflect the overall performance of different models through the final score, but also effectively distinguish performance differences across different task types such as open scenes, object relations, and text generation, thereby providing a more interpretable basis for model comparison and failure case analysis.

D. COMPARATIVE EXPERIMENTS WITH EXISTING AUTOMATIC EVALUATION METHODS

To further verify the rationality of the proposed five-dimensional evaluation framework in its corresponding dimensions, this paper selects existing automatic evaluation methods as nearby baselines for comparison. Specifically, we align existing methods with the corresponding dimensions of our framework: CLIPScore [18] for Global Semantic Alignment, OCR-based Text Fidelity for Text Fidelity, and ImageReward [14] as an overall baseline for comparison with our Final Score. It should be noted that although these baselines are similar to the proposed method in evaluation objectives, they are not exactly the same in underlying implementation, scoring scale, and post-processing. Therefore, this paper focuses more on their consistency and complementarity at the trend level, rather than on a simple comparison of absolute numerical superiority or inferiority.

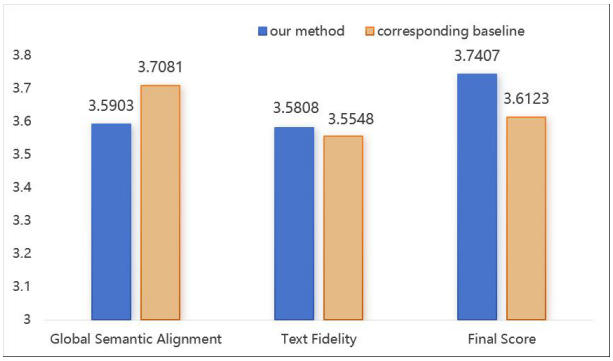

FIGURE 5. Comparison results between the proposed method and corresponding baselines under a unified five-point scale. Each group compares the average scores of the proposed Global Semantic Alignment, Text Fidelity, and Final Score with their corresponding baselines, namely CLIPScore, OCR-based Text Fidelity, and ImageReward.

The results in Fig. 5 indicate that our method achieves a Text Fidelity score of 3.5808, 0.0260 more than OCR-based Text Fidelity's score 3.5548. It means that the proposed method is still highly consistent with the existing OCR-based evaluation when assessing explicit text consistency but yields a slightly higher score. In terms of global semantic evaluation, the Global Semantic Alignment score yielded by our method is 3.5903. In comparison, CLIPScore obtains a score of 3.7081. Although the scores are quite close, both methods share a similar evaluation trend to some degree. As for the overall score, we yield a Final Score of 3.7407, while mapping ImageReward output into a five-point scale can obtain a score of 3.6123. In summary, all the above comparisons have confirmed that the proposed Text Fidelity score is very close in numerical value to OCR-based Text Fidelity, which indicates the consistency of our text dimension with OCR-driven evaluation metrics. Moreover, our final score is quite close to ImageReward. On the other hand, a more significant difference was found between our Global Semantic Alignment score and CLIPScore. Both methods utilize CLIP for semantic evaluation. However, the discrepancy should be caused by systematic differences in their normalization, score mapping, and post-processing procedures.

E. DISCUSSION OF RESULTS

From the comprehensive quantitative tables, SD-XL parameter sensitivity experiments, automatic baseline comparisons, visual cases, and comparison experiments with similar indicators, it can be seen that the proposed method has four major advantages. First, the scores of the five dimensions help researchers quickly locate the sources of errors. For example, when the total score of an image is not high, one can further determine whether the issue is text recognition failure, object omission, relational error, perceptual quality degradation, or overall semantic drift. Second, the method can be used both for horizontal comparison across different models and for parameter tuning analysis of the same model under different sampling configurations, and therefore has good engineering application value. Third, the multi-dimensional evaluation results are highly consistent with visual cases: high-scoring samples usually perform more stably in object presence, relational satisfaction, and global semantic matching, whereas low-scoring samples often have obvious shortcomings in one or more local dimensions. Finally, through score comparisons with other similar evaluation systems, the experimental results indicate that the proposed method can effectively distinguish the overall performance of different models and is relatively accurate.

V. CONCLUSION

This paper proposes a multi-dimensional image quality assessment method for text-to-image generation, constructing a unified scoring framework from five dimensions: OCR-based text fidelity, perceptual quality, object consistency, relational consistency, and global semantic alignment. Based on updated multi-model experimental results, SD-XL parameter sensitivity analysis, automatic baseline comparisons, and representative visual cases, this paper verifies the effectiveness of the proposed method in distinguishing model performance, explaining score sources, and assisting result diagnosis. Compared with a single automatic metric, the proposed method is more suitable for single-image diagnosis, horizontal model comparison, and sampling parameter tuning. In the future, we plan to consider human evaluation and improved models in the vision-language area, as well as more detailed benchmarks for text rendering, so that we can achieve higher reliability in determining weights and higher interpretability of the score. The proposed approach will be applied to difficult cases, such as multi-object compositions and fine-grained attributes.

DISCLOSURES

The authors state that there is no conflict of interest of any kind in the form of financial gain or professional associations that might have compromised their research in its formulation or execution.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 27, 2014.

[2] R. Rombach, A. Blattmann, D. Lorenz, et al., "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695, doi: 10.1109/CVPR52688.2022.01042.

[3] A. Ramesh, M. Pavlov, G. Goh, et al., "Zero-shot text-to-image generation," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8821–8831, doi: 10.48550/arXiv.2102.12092.

[4] C. Saharia, W. Chan, S. Saxena, et al., "Photorealistic text-to-image diffusion models with deep language understanding," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2022, doi: 10.48550/arXiv.2205.11487.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[6] T. Salimans, I. Goodfellow, W. Zaremba, et al., "Improved techniques for training GANs," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2016, doi: 10.48550/arXiv.1606.03498.

[7] M. Heusel, H. Ramsauer, T. Unterthiner, et al., "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, doi: 10.48550/arXiv.1706.08500.
[8] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763, doi: 10.48550/arXiv.2103.00020.
[9] J. Li, D. Li, C. Xiong, et al., "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. Int. Conf. Mach. Learn. (ICML), 2022, pp. 12888–12900, doi: 10.48550/arXiv.2201.12086.
[10] J. Li, D. Li, S. Savarese, et al., "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in Proc. Int. Conf. Mach. Learn. (ICML), 2023, pp. 19730–19742, doi: 10.48550/arXiv.2301.12597.
[11] Y. Hu, Z. Luo, D. Zhang, et al., "TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 20413–20422, doi: 10.48550/arXiv.2307.06350.
[12] K. Huang, K. Sun, E. Xie, et al., "T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2023, doi: 10.48550/arXiv.2307.06350.
[13] D. Ghosh, A. Gupta, and Y. Bengio, "GenEval: An object-focused framework for evaluating text-to-image alignment," in Adv. Neural Inf. Process. Syst. Datasets Benchmarks Track (NeurIPS), 2023, doi: 10.48550/arXiv.2310.11513.
[14] J. Xu, X. Liu, Y. Wu, et al., "ImageReward: Learning and evaluating human preferences for text-to-image generation," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2023, doi: 10.48550/arXiv.2304.05977.
[15] Y. Kirstain, A. Polyak, U. Singer, et al., "Pick-a-Pic: An open dataset of user preferences for text-to-image generation," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2023, doi: 10.48550/arXiv.2305.01569.
[16] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018, doi: 10.1109/TIP.2018.2831899.
[17] J. Ke, Q. Wang, Y. Wang, et al., "MUSIQ: Multi-scale image quality transformer," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 5148–5157, doi: 10.1109/ICCV48922.2021.00510.
[18] J. Hessel, A. Holtzman, M. Forbes, et al., "CLIPScore: A reference-free evaluation metric for image captioning," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2021, pp. 7514–7528, doi: 10.18653/v1/2021.emnlp-main.595.
[19] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, Jun. 2008, doi: 10.1049/el:20080522.
[20] J. Redmon, S. Divvala, R. Girshick, et al., "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.