

Q-Eval-100K: Evaluating Visual Quality and Alignment Level for Text-to-Vision Content

Zicheng Zhang¹ Tengchuan Kou¹ Shushi Wang¹ Chunyi Li¹
 Wei Sun¹ Wei Wang² Xiaoyu Li² Zongyu Wang² Xuezhi Cao²
 Xionghuo Min¹ Xiaohong Liu^{1†} Guangtao Zhai¹
¹Shanghai Jiao Tong University ²Meituan

Abstract

Evaluating text-to-vision content hinges on two crucial aspects: visual quality and alignment. While significant progress has been made in developing objective models to assess these dimensions, the performance of such models heavily relies on the scale and quality of human annotations. According to **Scaling Law**, increasing the number of human-labeled instances follows a predictable pattern that enhances the performance of evaluation models. Therefore, we introduce a comprehensive dataset designed to Evaluate Visual quality and Alignment Level for text-to-vision content (**Q-EVAL-100K**), featuring the largest collection of human-labeled Mean Opinion Scores (MOS) for the mentioned two aspects. The **Q-EVAL-100K** dataset encompasses both text-to-image and text-to-video models, with 960K human annotations specifically focused on visual quality and alignment for 100K instances (60K images and 40K videos). Leveraging this dataset with context prompt, we propose **Q-Eval-Score**, a unified model capable of evaluating both visual quality and alignment with special improvements for handling long-text prompt alignment. Experimental results indicate that the proposed **Q-Eval-Score** achieves superior performance on both visual quality and alignment, with strong generalization capabilities across other benchmarks. These findings highlight the significant value of the **Q-EVAL-100K** dataset. Data and codes will be available at <https://github.com/zzc-1998/Q-Eval>.

1. Introduction

With the rapid advancement of generative AI, millions of text-to-image and text-to-video content are being generated daily across various platforms [26, 57], applied in industrial production or directly used by consumers. However, due to current technological limitations, text-to-vision con-

[†] Corresponding author.

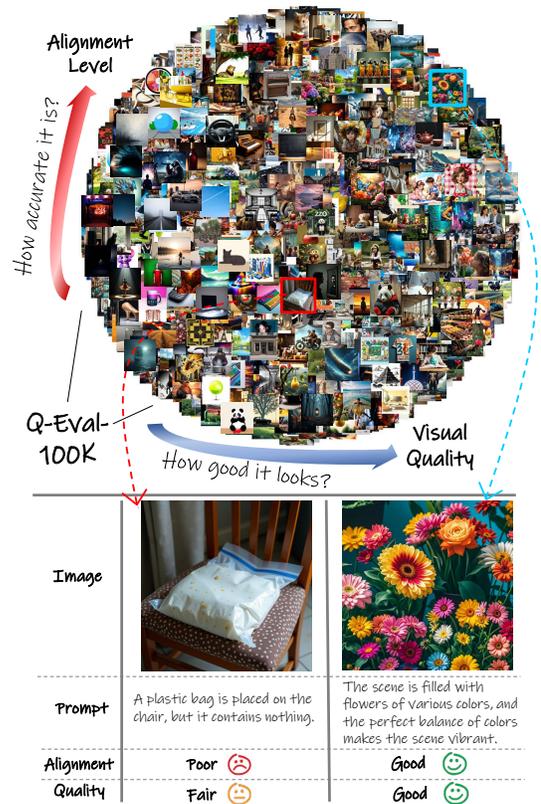


Figure 1. Illustration of the unified evaluation dimensions of **Q-Eval-100K**. We focus on **visual quality** (including all factors that may impact the viewing experience) and **alignment level**, which measures the accuracy of the generated content to the prompt.

tent often falls short of being perfect upon generation and cannot be immediately deployed [35, 37, 41], which usually requires expert evaluation, editing, and fine-tuning. As a result, numerous efforts have been made to develop automated methods for evaluating text-to-vision content, aiming to control the quality of generation and guide the necessary optimizations effectively [14, 27, 28, 30, 41, 61–64, 66, 71, 73]. Through extensive theoretical and experimental analy-

Table 1. A brief comparison of the latest text-to-vision evaluation datasets (*I.* for image, *V.* for video). For **Annotation Type**, **SBS** (*side-by-side*) and **MOS** (*mean-opinion-score*) refer to selecting the preferred instance from a pair of instances and assigning an average absolute score to a single instance respectively. For **Rating Concerns**, *Overall* indicates assigning scores from a holistic perspective while \checkmark denotes assigning separate scores to quality or alignment. For **Number**, *Ins.* and *Ann.* stand for the number of instances and human annotations respectively.

Dataset	Year	Content	Annotation Type (Single/Pair)	Evaluation Concern		Number Ins./Ann.
				Quality	Alignment	
Pick-a-pic [28]	2023	I.	SBS	Overall		1M/500K
ImageReward [66]	2023	I.	SBS	Overall		68k/137k
HPDv2 [64]	2023	I.	SBS	Overall		430k/645K
AGIQA-3k [35]	2023	I.	MOS	\checkmark	\checkmark	3K/81K
AIGCIQA2023 [59]	2023	I.	MOS	\checkmark	\checkmark	2K/17K
PKU-AIGIQA-4k [68]	2024	I.	MOS	\checkmark	\checkmark	4K/84K
AIGIQA-20k [36]	2024	I.	MOS	Overall		20K/420K
RichHF [40]	2024	I.	MOS	\checkmark	\checkmark	18K/54K
VideoFeedback [20]	2024	V.	MOS	\checkmark	\checkmark	37.6K/37.6K
T2VQA-DB [29]	2024	V.	MOS	Overall		10K/27K
GenAI-Bench [34]	2024	I.&V.	1-5 Likert Scale	Overall		9.6K/9.6K
Q-Eval-100K (Ours)	2024	I.&V.	MOS	\checkmark	\checkmark	100K/960K

sis, the evaluation of text-to-vision content can be primarily divided into two dimensions [42, 65, 66, 74, 77]: **Visual Quality** (the perceived quality of the visual content, which can be simply understood as *how good it looks*) and **Alignment** level (the consistency between text and vision, which can be interpreted as *how accurate the generation is*).

To meet the need for evaluation, many text-to-vision evaluation datasets have been proposed, along with corresponding evaluation algorithms [12, 13, 20, 20, 25, 29, 34–36, 40, 42, 59, 68, 72, 75, 76]. However, these efforts face the following significant limitations: 1) **The key evaluation dimensions for text-to-vision content are often not systematically captured.** Some datasets propose too many dimensions, adding unnecessary complexity to the evaluation process. The practical applicability of these dimensions can be narrow or lead to redundancy. 2) **Most text-to-vision evaluation datasets fail to disentangle visual quality and alignment.** These datasets either focus solely on alignment or visual quality, or merge both dimensions into a single score, leading to results that are often incomplete and ambiguous, making it challenging to address specific evaluation needs. 3) **The scale of these datasets remains insufficient.** With the rise of Large Multimodal Models (LMMs), which have demonstrated strong capabilities in visual and textual understanding, researchers are increasingly leveraging them for text-to-vision evaluation. However, current dataset sizes remain inadequate to fully unlock the potential of LMM-based models [77], conceivably restricting their applicability and generalization in real-world scenarios.

To address these challenges, we present **Q-Eval-100K**, which, to the best of our knowledge, is the largest text-to-vision evaluation dataset with Mean Opinion Scores (MOSs), comprising 100K instances (including 60K generated images & 40K generated videos). A brief comparison of **Q-Eval-100K** and previous text-to-vision eval-

uation datasets is illustrated in Table 1. We manually gather prompts from existing benchmarks and create diverse prompts that focus on three key aspects: *entity generation*, *entity attribute generation*, and *interaction capability*. The instances in **Q-Eval-100K** are then generated from a diverse range of generative models, both open-source and closed-source, to ensure high diversity. We implement a rigorous, scientifically grounded subjective evaluation process using a **Sample & Scrutinize** strategy, focusing on both visual quality and alignment level for each of the 100K instances, yielding a total of high-quality 960K human annotations.

Building on the proposed **Q-Eval-100K**, we propose **Q-Eval-Score**, a unified evaluation framework capable of assessing both visual quality and alignment, providing separate scores for each dimension. We first adapt **Q-Eval-100K** into a Supervised Fine-Tuning (SFT) dataset optimized for injecting knowledge into LMMs. Scores are transformed into *adjective-based ratings*, then reformulated within a well-guided *context-prompt format*. Specifically, for visual quality, we guide the model to identify positive and negative visual impacts, evaluate the intensity of these impacts, and make a balanced judgment. For alignment, we guide the model to perceive the overall situation, examine details, and reach a balanced judgment. The fine-tuning process is then supervised by a combined CE and MSE loss. During inference, the final score is computed as a weighted average based on *the probability of each rating token*. Notably, in handling long-prompt alignment, we observed that direct alignment assessment often yields low scores due to oversimplification. To address this, we propose a **Vague-to-Specific** strategy, where a long prompt is converted into a vague version retaining only core information and multiple prompts with specific details. These prompts are evaluated separately and the alignment scores are combined to the final score. Our contributions can be summarized as follows:

- We present **Q-Eval-100K**, the largest text-to-vision evaluation dataset with MOSs, comprising 100K instances from various generative models. We employ a scientifically grounded evaluation methodology, using a **Sample & Scrutinize** strategy to collect 960K human annotations focusing on visual quality and alignment.
- We propose **Q-Eval-Score**, a unified evaluation framework capable of independently assessing visual quality and alignment, providing separate scores for each dimension. Specifically, we adapt **Q-Eval-100K** into an SFT dataset with adjective-based ratings in a structured context-prompt format for enhancing the visual quality and alignment evaluation capabilities of LMMs.
- To improve alignment evaluation for long prompts, we introduce a **Vague-to-Specific** strategy, which separates prompts into core and detailed variants, yielding a more accurate alignment score through weighted averaging.

2. Related Works

2.1. Benchmarks for Text-to-Vision Evaluation

Early text-to-vision benchmarks largely depend on multi-modal datasets labeled with captions [11, 23, 44]. However, with increasing recognition of human feedback’s value [15, 24, 77], many benchmarks begin to employ human annotations. Common annotation methods include SBS (side-by-side) and MOS (mean opinion score) [50, 69]. SBS requires selecting a preferred instance from a pair, while MOS assigns a score to a single instance. SBS is generally easier for human subjects and more precise, but MOS is more versatile and broadly applicable to various situations [7, 33].

Text-to-vision evaluation dimensions [42, 74] can be categorized into visual quality and alignment. While some benchmarks [12, 13, 20, 25] treat aspects like naturalness, aesthetics, and temporal consistency as distinct dimensions, we view these as components of visual quality since they collectively influence the quality of experience (QoE) for viewers. Early benchmarks [35, 36, 59, 68] for generated images comprehensively address both visual quality and alignment for evaluation. RichHF [40] enhances these evaluations by incorporating subjective scores, heatmaps, and misalignment tokens. For video, VideoFeedback [20] introduces five dimensions of quality and alignment, while T2VQA-DB [29] focuses primarily on visual quality. Further, GenAI-Bench [34] evaluates alignment for both generated images and generated videos. The proposed **Q-Eval-100K** dataset offers a unified text-to-vision evaluation framework, significantly increasing dataset scale and diversity, making it distinct from prior benchmarks.

2.2. Metric for Text-to-Vision Evaluation

Previous evaluation methods separately focus on either visual quality or alignment. Perceptual methods assess the visual quality of generated content, utilizing traditional scores like IS [53], FID [22], and LPIPS [70] with pre-trained neural networks. Recently, data-driven models [35] trained on specialized datasets have further advanced perceptual score prediction. Additionally, methods such as CLIP-IQA [58] and Q-Align [63] leverage text prompts to enhance perceptual alignment. Alignment methods, integrating both text and vision modalities, initially use CLIP-Score [21] due to its ease of application. To address more complex prompts, some approaches [28, 64, 66] incorporate human feedback to improve evaluation accuracy. Given the powerful interpretative capabilities of LMMs, recent work [14, 27, 30, 41, 71] has begun to apply these models to alignment assessments. Most existing models evaluate either perceptual quality or alignment exclusively. The proposed **Q-Eval-Score** addresses this gap by offering decoupled scores for both perceptual quality and alignment.

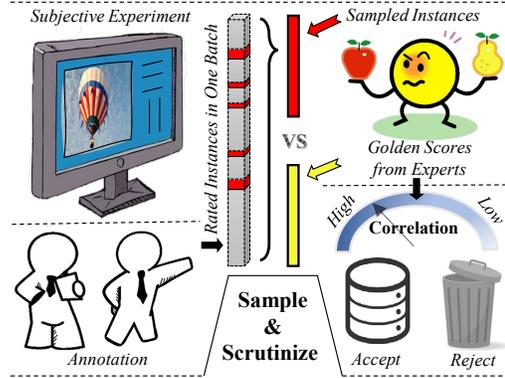


Figure 2. Illustration of the **Sample and Scrutinize** quality control strategy for annotations in **Q-Eval-100K**. We randomly select a sample of 5K instances from the full dataset, which are then reviewed by experts to establish golden scores. A batch of annotations is approved only if the scores of the sampled instances show a high correlation with these expert-assigned golden scores.

3. Q-Eval-100K Construction

3.1. Basic Principles

The construction process of **Q-Eval-100K** is illustrated in Fig. 3. We follow these guiding principles: 1) **Ensuring diversity** in generated content by collecting a wide range of prompts and using multiple generative models; 2) **Ensuring annotation quality** through carefully designed experimental settings and standards to achieve the highest accuracy; 3) **Ensuring effective learning**, through adapting the data for LMM suitability by transforming both visual quality and alignment scores into a context-aware SFT dataset.

3.2. Sources Collection

Prompt Designing. The prompt design focuses on three main aspects: **Entity Generation**, **Entity Attribute Generation**, and **Interaction Capability**. 1) Entity generation targets the primary entities (*people, objects, etc.*) to be generated. 2) Entity attribute generation emphasizes the attributes (*clothing, color, material, etc.*) of the entities. 3) Interaction capability focuses on the interactions between the generated entities and other entities or the background, such as their spatial relationships and actions. Following the outlines mentioned above, we manually create a portion of the prompts and extract some from existing datasets [34, 47].

Generation Models. We utilize multiple popular text-to-image and text-to-video models to ensure diversity, which include FLUX [31], Lumina-T2X [18], PixArt [10], Stable Diffusion 3 [4], Stable Diffusion XL [51], DALL-E 3 [48], Wanx [16], Midjourney [45], Hunyuan-DiT [39], Kolors [55], ERNIE-ViLG [17], CogVideoX [67], Runway GEN-2 [19], Runway GEN-3 [52], Latte [43], Kling [54], Dreamina [9], Luma [2], PixVerse [3], Pika [32], Stable Video Diffusion [8], Vidu [56].

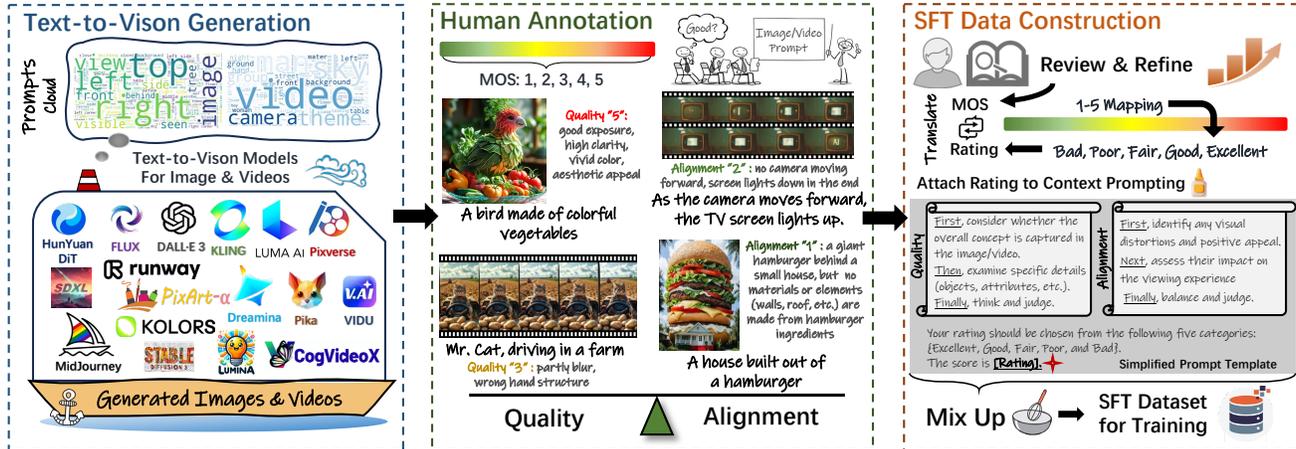


Figure 3. Overview of the **Q-Eval-100K** construction process. We design a wide range of prompts and employ various text-to-visual models to generate diverse content. Subjective evaluations are then conducted to rate the visual quality and alignment of these generated instances. The resulting scores form the SFT dataset, which can help inject corresponding knowledge into LLMs.

3.3. Subjective Experiment

Given the large scale of **Q-Eval-100K**, we develop rigorous experimental protocols to ensure the highest possible accuracy in our annotations. To facilitate this, we establish a well-controlled indoor experimental environment, where more than 200 human subjects are recruited to participate in the annotation. To ensure the accuracy of annotations is not compromised by individual cognitive differences or annotator fatigue, we propose a **Sample & Scrutinize** data control strategy as shown in Fig. 2. The strategy includes two steps: 1) First, we randomly sample 5,000 instances from the entire dataset. We then organize experts with rich experience to discuss and score these instances, leading to the establishment of golden scores for both visual quality and alignment. These golden scores remain hidden for all subsequent experimental subjects. 2) Next, we provide human annotators with comprehensive training before they begin the annotation process. After each batch, we gather the scores for instances that have golden scores and compare them with these golden scores, calculating the correlation values (SRCC-rank similarity). Only batches with an SRCC above 0.8 are accepted, otherwise, they are rejected.

Additionally, we split **Q-Eval-100K** into training and testing sets in an 80:20 ratio. Each instance in the training set has at least three annotations, while each instance in the testing set has a minimum of twelve annotations to ensure accuracy. This process results in a total of over 960K annotations, calculated as follows: $80K$ (training instances) \times 2 (visual quality & alignment) \times 3 (minimum annotation number) + $20K$ (training instances) \times 2 (visual quality & alignment) \times 12 (minimum annotation number) = 960K annotations. Finally, we calculate the average of the multiple annotations to derive the score for each instance.

3.4. Statistical Analysis

The distributions of MOSs for visual quality and alignment are exhibited in Fig. 4 respectively, which reveal several key insights. In general, **there are substantial differences among generation models in both visual quality and alignment**, with their distributions displaying significant inconsistencies, indicating varied performance across different generation prompts. 1) For image alignment, distributions are generally skewed higher, suggesting that models perform relatively well in aligning images, though multiple peaks in the 4-5 and 2-3 score ranges indicate some fluctuation in performance. 2) In video alignment, model performance varies more markedly, with most scores concentrated between 2 and 4, highlighting the need for improvement in alignment in video generation. 3) Visual quality for images scores noticeably lower than image alignment, indicating that generation models perform significantly worse in visual quality. Furthermore, the wider distribution spread in image visual quality suggests greater variance and instability across models. 4) Similarly, video visual quality scores are lower than alignment scores, highlighting a consistent underperformance in visual quality. Interestingly, models such as Kling, Dreamina, Luma, PixVerse, and Pika exhibit similar distributions for alignment and visual quality, indicating consistent capability across both aspects. However, this consistency is not observed across all models.

Overall, the findings above highlight a notable disparity in that visual quality generally falls behind alignment. This gap likely stems from the current emphasis on alignment optimization, which is also relatively easier to improve, whereas visual quality has received less focus. This analysis underscores the importance of **Q-Eval-100K** as a comprehensive benchmark for evaluating both dimensions.

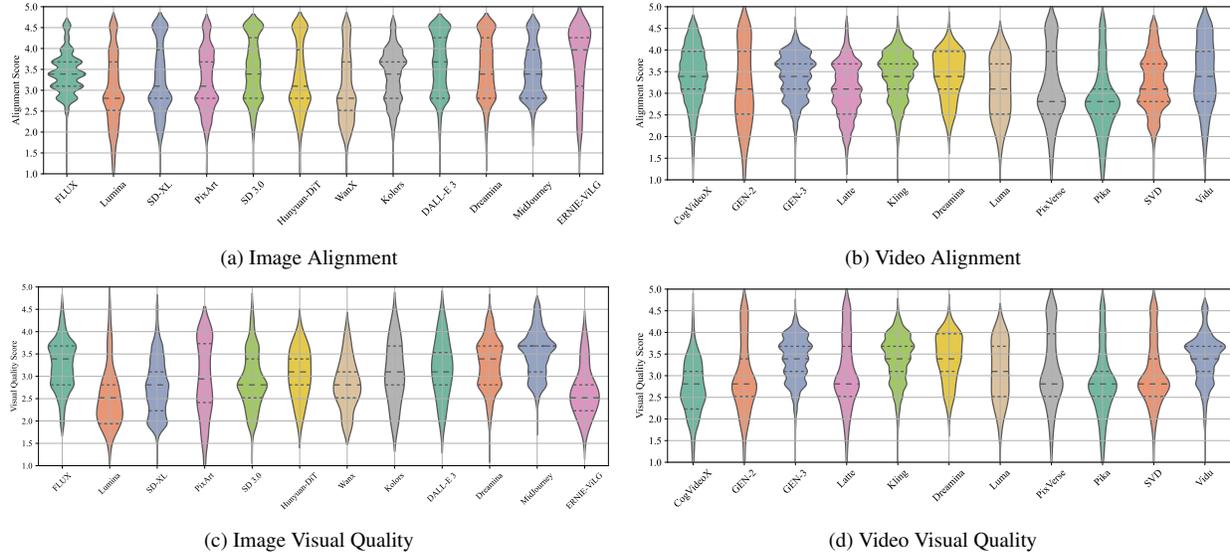


Figure 4. MOS distributions for the visual quality and alignment of generated images and videos in the **Q-Eval-100K** dataset respectively.

4. Q-Eval-Score

4.1. Unified Pipeline for Decoupled Evaluation

Although the evaluation of visual quality and alignment are two relatively independent tasks, we leverage the adaptability and extensive prior knowledge of LMMs to propose a unified model, **Q-Eval-Score**, that addresses both visual quality and alignment level evaluation within a single framework. Specifically, we convert the human-labeled MOS from the **Q-Eval-100K** dataset for both visual quality and alignment levels into a fixed-prompt format, creating a mixed SFT dataset. We then fine-tune the LMM, enabling it to evaluate both visual quality and alignment levels.

4.2. How to Teach LMMs to Evaluate

4.2.1 Context Prompt

In previous work using prompts with LMMs for evaluation, the questions are often straightforward and simple, such as ‘Can you evaluate the quality of the image?’ (Q-Align [63]) or ‘Does this figure show [Prompt]? Please answer yes or no.’ (VQAScore [41]). However, this simplicity may lead to confusion for the model, as the prompts may not be specific enough to guide a more detailed or accurate evaluation.

Inspired by the chain-of-thought (CoT [60]) concept and given that humans undergo a reasoning process when evaluating visual quality and alignment, we propose a **Context-Prompt** format to construct our SFT dataset. For the visual quality task, the human evaluation process can be summarized as first identifying both positive and negative quality factors, then measuring these factors, and finally weighing them to reach a conclusion. Based on this process, we design the following prompt structure:

Context Prompt for Visual Quality

User: Suppose you are an expert in evaluating the visual quality of AI-generated image/video. First, identify any visual distortions and positive visual appeal regarding low-level features and aesthetics. Next, assess the severity of distortions and their impact on the viewing experience, noting whether they are subtle or distracting, and evaluate how the positive features enhance the visual appeal, considering their strength and contribution to the overall aesthetics. Finally, balance the identified distortions against the positive aspects and give your rating on the visual quality. Your rating should be chosen from the following five categories: [Excellent, Good, Fair, Poor, and Bad]. For this image/video [Image/Frames], the text prompt is [Prompt].
Answer: [Rating] (Excellent, Good, Fair, Poor, Bad).

For the alignment task, the human evaluation process involves observing whether the overall content generally aligns with the text, followed by a more detailed comparison, and finally a comprehensive evaluation for conclusion:

Context Prompt for Visual Quality

User: Suppose you are an expert in evaluating alignment between the text prompt and the AI-generated image/video. Begin by considering whether the overall concept of the prompt is captured in the image/video. Then, examine the specific details, such as the presence of key objects, their attributes, and relationships. Check if the visual content accurately reflects these aspects. Finally, give your alignment rating considering both overall and detailed accuracy. Your rating should be chosen from the following five categories: [Excellent, Good, Fair, Poor, and Bad]. For this image/video [Image/Frames], the text prompt is [Prompt].
Answer: [Rating] (Excellent, Good, Fair, Poor, Bad).

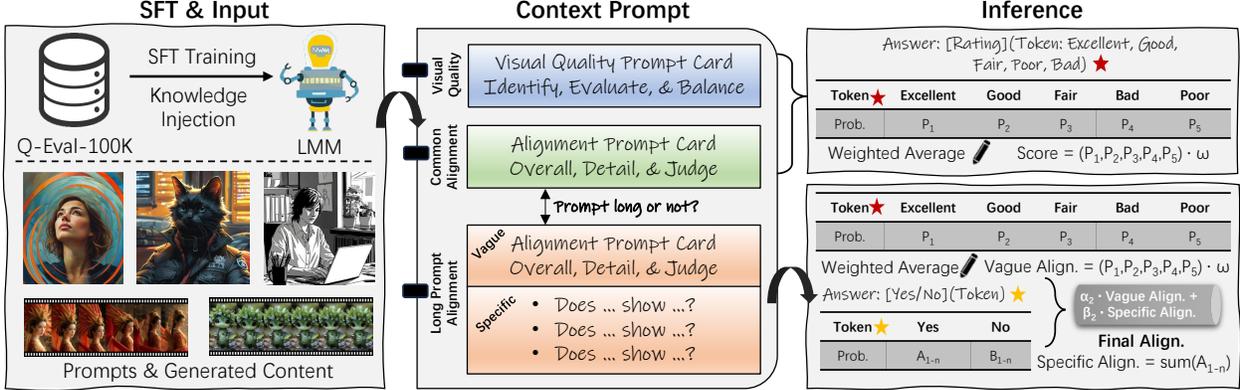


Figure 5. The pipeline of the proposed **Q-Eval-Score** model involves multiple stages. First, the **Q-Eval-100K** SFT dataset is used to train the LMM on visual quality and alignment knowledge. Then, context prompts are applied to guide the LMM towards generating more detailed and accurate outputs. Finally, the rating token probabilities are converted into predicted scores. Additionally, long prompt alignment is achieved through a **Vague-to-Specific** strategy to further refine the model’s responses.

4.2.2 ‘Translating’ MOS into Ratings

It is well established that discrete adjective ratings are easier for LMMs to interpret compared to numerical scores [63, 73]. Since MOS in **Q-Eval-100K** is labeled in absolute terms, we can easily map MOS to the corresponding rating:

$$R(s) = r_i \text{ if } m + \frac{i-1}{5} \times (M-m) < s \leq m + \frac{i}{5} \times (M-m), \quad (1)$$

$$\{r_i \mid i=1 \sim 5\} = \{Bad, Poor, Fair, Good, Excellent\},$$

where $m = 1$ and $M = 5$ (score range bound of **Q-Eval-100K**), $R(s)$ indicates the mapped rating of MOS value s .

4.3. Model Architecture

Using the constructed SFT dataset with the question-answer pairs as described, we select Qwen-VL [6] as the LMM model (*Qwen2-VL-7B-Instruct*) for training, which has demonstrated strong visual understanding capabilities for both images and videos. For video processing, each video is converted into a sequence of images at a rate of one frame per second, which is then fed into the model as the input. The scoring computation method is detailed as follows. For the rating token, we first calculate the model output probabilities p_j for each of the five rating terms $\{Excellent, Good, Fair, Poor, Bad\}$, where $j \in \{1, 2, 3, 4, 5\}$. Then we define the final predicted rating \hat{r} as the weighted average of these probabilities:

$$\hat{r} = \sum_{j=1}^5 p_j \cdot w_j, \quad (2)$$

where w_j is the numerical weight assigned to each rating (e.g., $w_j = \{1, 0.75, 0.5, 0.25, 0\}$ for *Excellent* to *Bad*).

4.4. Loss Function

The loss function for the model consists of two parts: Cross-Entropy (CE) Loss and Mean Squared Error (MSE) Loss. The CE Loss can assist the LMM in learning the general question-answer format and necessary knowledge. Meanwhile, the MSE Loss refines the score prediction accuracy. The CE Loss for question-answer pairs is defined as:

$$\mathcal{L}_{CE} = - \sum_{i=1}^N y_i \cdot \log(p_i), \quad (3)$$

where y_i is the one-hot encoded vector representing the true label for instance i , and p_i is the predicted probability vector for the answer tokens. The MSE Loss can then be given by:

$$\mathcal{L}_{MSE} = (\hat{r} - r_{MOS})^2, \quad (4)$$

where \hat{r} and r_{MOS} represent the predicted scores and the MOS labels respectively. The total loss \mathcal{L} is a weighted sum of the CE Loss and MSE Loss:

$$\mathcal{L} = \alpha_1 \cdot \mathcal{L}_{CE} + \beta_1 \cdot \mathcal{L}_{MSE}, \quad (5)$$

where α_1 and β_1 (default set as 1 & 1) are weight parameters controlling the contribution of each loss term.

4.5. Handling Long Prompt Alignment

During inference, we observe that the trained LMM tends to undervalue alignment when handling long prompts (more than 25 words). This is partly because long prompts are underrepresented in the training data, leading to insufficient training. More importantly, the LMM acts as a strict evaluator, often penalizing significant points for inconsistencies that may seem minor to humans. These small discrepancies occur more frequently with long prompts. To manage this

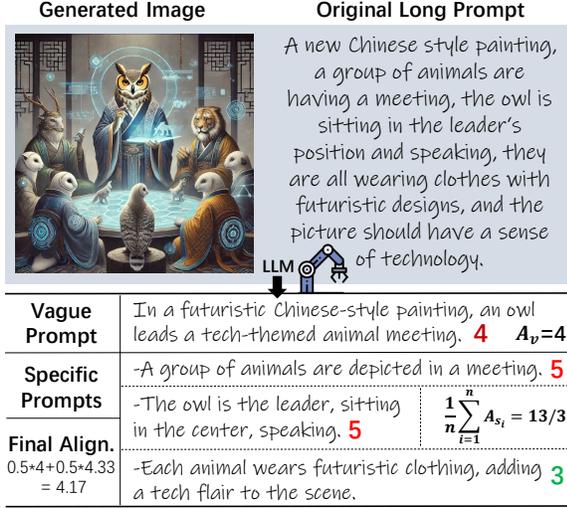


Figure 6. An example of the **Vague-to-Specific** strategy. The original long prompt is divided by the LLM (QwenLM [5]) into a **Vague Prompt** and several **Specific Prompts**. The alignment score is first calculated separately for each part, then combined using weighted averaging to form the final score.

issue, we propose a **Vague-to-Specific** strategy. We use an additional LLM (QwenLM [5]) to summarize long prompts, retaining only the core features while filtering out details, producing a concise **Vague Prompt**. Then, we split the long prompt into **Specific Prompts** (*no more than three*), each maintaining full details but avoiding redundancy:

$$(P_v, \{P_{s_1}, \dots, P_{s_n}\}) = \mathcal{VS}(P_{Long}), \quad (6)$$

where P_v represents the **Vague Prompt**, $\{P_{s_1}, \dots, P_{s_n}\}$ stands for the set of **Specific Prompts**, $\mathcal{VS}(\cdot)$ indicates the prompt split function, and P_{Long} is the original long prompt. For the **Vague Prompt**, we calculate alignment in the usual way. However, directly asking for consistency with the **Specific Prompts** is not appropriate since each one addresses only part of the vision content. Drawing inspiration from the VQAScore [41] approach, we modify the question to a softer format, such as ‘Does the image/video show [Prompt]?’ to evaluate alignment (measuring as the logit probability of answering ‘Yes’) for each **Specific Prompts**. Finally, we combine the results from both the **Vague Prompt** and **Specific Prompt** using a weighted approach to calculate the final alignment score:

$$\mathcal{A}_f = \alpha_2 \mathcal{A}_v + \beta_2 \left(\frac{1}{n} \sum_{i=1}^n \mathcal{A}_{s_i} \right), \quad (7)$$

where \mathcal{A}_f , \mathcal{A}_v , and \mathcal{A}_{s_i} are the alignment scores for the final evaluation, vague prompt, and i -th specific prompt. α_2 and β_2 (0.5 & 0.5 as default) are weight parameters.

Table 2. Performance comparison on the visual quality aspect of **Q-Eval-100K**. Best in **bold**, second underlined.

Model (Visual Quality)	Q-Eval-Image				Q-Eval-Video			
	Instance-level		Model-level		Instance-level		Model-level	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
NIQE	0.239	0.238	0.829	0.758	-0.057	-0.074	-0.333	-0.241
CLIP-IQA	0.334	0.324	0.600	0.685	0.194	0.175	0.095	0.076
Q-Align	<u>0.587</u>	<u>0.578</u>	0.714	0.914	<u>0.500</u>	<u>0.502</u>	0.762	<u>0.762</u>
IPCE	0.550	0.560	<u>0.933</u>	<u>0.937</u>	0.299	0.302	0.476	0.568
Q-Eval-Score	0.732	0.731	0.943	0.949	0.601	0.609	0.762	0.814

Table 3. Performance comparison on the alignment aspect of **Q-Eval-100K**. Considering that CLIPScore, BLIP2Score, and VQAScore are popular zero-shot alignment evaluation metrics, we provide the corresponding performance with the official default weight as well (marked with *).

Model (Align- ment)	Q-Eval-Image				Q-Eval-Video			
	Instance-level		Model-level		Instance-level		Model-level	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
CLIPScore*	0.245	0.252	0.617	0.685	0.186	0.219	0.518	0.500
BLIP2Score*	0.297	0.330	0.764	0.835	0.218	0.250	0.295	0.296
VQAScore*	0.549	0.468	0.385	0.555	0.433	0.432	0.433	0.351
CLIPScore	<u>0.768</u>	<u>0.740</u>	<u>0.958</u>	<u>0.956</u>	0.431	0.443	<u>0.519</u>	<u>0.509</u>
BLIP2Score	0.766	<u>0.743</u>	0.933	0.934	<u>0.483</u>	<u>0.488</u>	0.512	0.481
ImageReward	0.762	0.732	0.925	0.955	0.472	0.485	0.375	0.362
Q-Eval-Score	0.822	0.802	0.964	0.969	0.607	0.634	0.648	0.605

5. Experiment

5.1. Experimental Setup

Training & Evaluation. The Qwen2-VL-7B-Instruct [6] serves as the backbone LMM for **Q-Eval-Score**. All visual quality and alignment data from images and videos are combined for training. Training is conducted on 8 A100 GPUs for one epoch by default. For evaluation metrics, we use SRCC and PLCC, which measure the rank and linear correlation between predicted scores and MOSs. We propose evaluations at the **Instance-level** and **Model-level** which assess accuracy in ranking specific generated instances and generative models based on overall performance.

Competitors. Few models can simultaneously predict both visual quality and alignment. Thus we selected task-specific competitors for each sub-task: For visual quality, we include NIQE [46], CLIP-IQA [58], Q-Align [63], and IPCE [49] (the top method from the ‘NTIRE 2024 Quality Assessment of AI-Generated Content Challenge’ [42]). For alignment, we choose CLIPScore [21], BLIP2Score [38], ImageReward [66] and VQAScore [41] as the competitors. All models are trained and tested using their default recommended parameters and the corresponding train-test sets of the **Q-Eval-100K** dataset unless specified.

5.2. Discussion & General Findings

The general performance on the visual quality and alignment is exhibited in Table 2 and Table 3, from which we can draw several conclusions: 1) For visual quality, The proposed **Q-Eval-Score** outperforms all competitors, achiev-

Table 4. Ablation Study of Q-Eval-Score.

Model	Q-Eval-Image (Quality)				Q-Eval-Video (Quality)				Q-Eval-Image (Alignment)				Q-Eval-Video (Alignment)			
	Instance-level		Model-level		Instance-level		Model-level		Instance-level		Model-level		Instance-level		Model-level	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
w/o SFT Training	0.071	0.096	0.257	0.136	0.018	0.008	0.262	0.314	0.529	0.423	0.560	0.705	0.464	0.437	0.567	0.478
w/o Context Prompt	0.504	0.509	0.600	0.756	0.598	0.591	0.571	0.638	0.805	0.776	0.960	0.963	0.588	0.597	0.601	0.602
w/o CE Loss	0.652	0.622	0.932	0.910	0.247	0.249	0.071	0.239	0.804	0.776	0.948	0.961	0.604	0.626	0.642	0.593
w/o MSE Loss	0.665	0.673	0.933	0.941	0.595	0.583	0.690	0.712	0.795	0.763	0.954	0.958	0.580	0.605	0.624	0.599
Q-Eval-Score	0.732	0.731	0.943	0.949	0.601	0.609	0.762	0.814	0.822	0.802	0.964	0.969	0.607	0.634	0.648	0.605

Table 5. Performance comparison on the alignment aspect of Q-Eval-100K on the long prompt subset, where w/o V2S and w V2S represents the proposed Q-Eval-Score model with and without the Vague-to-Specific strategy respectively.

Model (Alignment)	Q-Eval-Image (Long)		Q-Eval-Video (Long)	
	Instance-level		Instance-level	
	SRCC	PLCC	SRCC	PLCC
CLIPScore	0.533	0.547	0.359	0.367
BLIP2Score	0.620	0.636	0.392	0.395
VQAScore	0.432	0.325	0.344	0.350
w/o V2S	0.591	0.599	0.480	0.470
w V2S	0.620	0.623	0.517	0.512

Table 6. Cross-dataset validation performance on GenAI-Bench. The Q-Eval-Score is trained on the Q-Eval-100K and then validated on GenAI-Bench. * indicates using default weight.

Model (Alignment)	GenAI-Bench (Image)		GenAI-Bench (Video)	
	SRCC	PLCC	SRCC	PLCC
CLIPScore*	0.174	0.169	0.269	0.269
BLIP2Score*	0.221	0.209	0.289	0.275
VQAScore*	0.556	0.502	0.527	0.505
CLIPScore	0.681	0.670	0.610	0.628
BLIP2Score	0.687	0.679	0.679	0.705
ImageReward	0.664	0.656	0.663	0.684
Q-Eval-Score	0.757	0.747	0.717	0.714

ing the best performance overall. The decline in video performance is likely due to the 1fps frame sampling method, which causes a loss of temporal information, leading to inaccurate estimations. Despite this, at the instance-level, Q-Eval-Score still leads the second-best competitor (Q-Align) by 10% on video instance-level SRCC. 2) For alignment, Q-Eval-Score also demonstrates a significant lead in alignment, outperforming competitors by 6% in image instance-level SRCC and 12% in video instance-level SRCC. Additionally, the substantial performance improvements seen in the trained competitors suggest that Q-Eval-100K provides valuable knowledge for alignment evaluation. 3) In comparison to alignment, Q-Eval-Score’s performance in visual quality is notably lower, indicating that predicting visual quality is more challenging. This is likely because alignment evaluation is more straightforward and objective, while visual quality perception is more complex and subjective, making it harder to assess. Overall, the proposed Q-Eval-Score exhibits exceptional potential in both visual quality and alignment, achieving over 0.94 performance at the image model-level, closely aligning with human evaluations. This strong performance not only highlights the robustness of the model but also underscores its promising ability to serve as an effective evaluation metric.

5.3. Further Experiments

I) Ablation Study. We conduct a detailed ablation study to assess the contribution of proposed strategies and CE/MSE loss. The results are presented in Table 4. It is clear that each of the strategies we proposed and both CE/MSE loss make a significant contribution to the final outcome.

II) Long Prompt. To test the Vague-to-Specific strategy for long prompt alignment, we select a subset of 5,000 instances from Q-Eval-100K that contain long prompts (over 25 words) for testing, performance shown in Table 5. Due to the limited data size, we present only the instance-level performance. The results clearly show that the Vague-to-Specific strategy significantly improves performance, indicating the effectiveness of handling long prompt alignment.

III) Cross-dataset Validation. To demonstrate the value of the Q-Eval-100K dataset, we conduct a cross-dataset validation, with performance results shown in Table 6. It is important to note that instances generated from prompts in GenAI-Bench are excluded from this validation. The results clearly show that models trained on Q-Eval-100K significantly outperform the current state-of-the-art VQAScore on GenAI-Bench by a large margin, providing strong evidence of the generalization value of the Q-Eval-100K dataset.

6. Conclusion

In conclusion, we introduce Q-Eval-100K, the largest text-to-vision evaluation dataset to date, featuring 100K instances and 960K human annotations for assessing visual quality and alignment. We also present Q-Eval-Score, a unified evaluation framework that leverages this dataset to provide separate scores for each dimension. Experimental results show that Q-Eval-Score outperforms existing methods, demonstrating its potential for more reliable, comprehensive assessments of text-to-vision models. Looking ahead, we hope this work can lay a strong foundation for further advancements in text-to-vision model promotion and real-world evaluation applications.

Acknowledgment

The work was supported in part by the National Natural Science Foundation of China under Grant 62301310, Grant 623B2073, and in part by Sichuan Science and Technology Program under Grant 2024NSFSC1426.

References

- [1] Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500, 2000. **3**
- [2] Luma AI. Dream machine: Ai video generator. <https://lumalabs.ai/dream-machine>, 2024. Accessed: 2024-11-03. **3**
- [3] PixVerse AI. Pixverse: Ai video creation platform. <https://pixverse.ai/>, 2024. Accessed: 2024-11-03. **3**
- [4] Stability AI. Stable diffusion 3.0. <https://stability.ai>, 2023. Available through the Stability AI Developer Platform. **3**
- [5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xi-aodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. **7**
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. **6, 7, 1**
- [7] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. **3**
- [8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. **3**
- [9] Dreamina by CapCut. Dreamina, 2023. Available online at <https://dreamina.capcut.com/>. **3**
- [10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. **3**
- [11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. **3**
- [12] Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Exploring the naturalness of ai-generated images. *arXiv preprint arXiv:2312.05476*, 2023. **2, 3**
- [13] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. *arXiv preprint arXiv:2406.06087*, 2024. **2, 3**
- [14] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023. **1, 3**
- [15] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. **3**
- [16] Alibaba Cloud. Wanx. <https://tongyi.aliyun.com/wanxiang/>, 2024. Accessed: 2024-11-07. **3**
- [17] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, and H. Wang. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023. **3**
- [18] Peng Gao, Le Zhuo, Chris Liu, , Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. **3**
- [19] Anastasis Germanidis. Gen-2: Generate novel videos with text, images or video clips. <https://runwayml.com/research/gen-2>, 2023. Accessed: 2024-11-07. **3**
- [20] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhui Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *ArXiv*, abs/2406.15252, 2024. **2, 3**
- [21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. **3, 7**
- [22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. **3**
- [23] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. **3**
- [24] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. **3**
- [25] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. **2, 3**

- [26] Everypixel Journal. Ai image statistics for 2024: How much content was created by ai. *Everypixel*, 2024. Accessed: 2024-10-14. **1**
- [27] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2023. **1, 3**
- [28] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. **1, 2, 3**
- [29] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024. **2, 3**
- [30] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. ViScore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023. **1, 3**
- [31] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2023. Accessed: 2024-10-22. **3**
- [32] Pika Labs. Pika: Ai video generation platform. <https://pika.art/>, 2024. Accessed: 2024-11-03. **3**
- [33] Lucie L ev eque, Meriem Outtas, Hantao Liu, and Lu Zhang. Comparative study of the methodologies used for subjective medical image quality assessment. *Physics in Medicine & Biology*, 66(15):15TR02, 2021. **3**
- [34] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. **2, 3, 1**
- [35] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An open database for ai-generated image quality assessment. *CoRR*, 2306.04717, 2023. **1, 2, 3**
- [36] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. Aigiqa-20k: A large database for ai-generated image quality assessment. *arXiv preprint arXiv:2404.03407*, 2(3):5, 2024. **2, 3**
- [37] Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Q-refine: A perceptual quality refiner for ai-generated image, 2024. **1**
- [38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. **7**
- [39] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchu Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. **3**
- [40] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19401–19411, 2024. **2, 3**
- [41] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. **1, 3, 5, 7**
- [42] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, et al. Ntire 2024 quality assessment of ai-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6337–6362, 2024. **2, 3, 7**
- [43] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. **3**
- [44] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. **3**
- [45] MidJourney. Midjourney. <https://www.midjourney.com>, 2023. Accessed: 2024-10-22. **3**
- [46] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. **7**
- [47] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docc: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*, 2024. **3, 1**
- [48] OpenAI. Dall-e 3. <https://openai.com/dall-e-3>, 2023. Accessed: 2024-10-22. **3**
- [49] Fei Peng, Huiyuan Fu, Anlong Ming, Chuanming Wang, Huadong Ma, Shuai He, Zifei Dou, and Shu Chen. Aige image quality assessment via image-prompt correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6432–6441, 2024. **7**
- [50] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019. **3**
- [51] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas M uller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. **3**

- [52] Runway. Introducing gen-3 alpha: A new frontier for video generation. <https://runwayml.com/research/introducing-gen-3-alpha>, 2024. Accessed: 2024-11-07. **3**
- [53] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. **3**
- [54] Kuaishou Team. Kling ai. <https://klingai.io/>, 2024. **3**
- [55] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024. **3**
- [56] Vidu AI Team. Vidu ai. <https://www.vidu.studio/zh>, 2024. Accessed: 2024-11-03. **3**
- [57] Techreport. Ai image generator market statistics in 2024. *Techreport*, 2024. Accessed: 2024-10-14. **1**
- [58] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images, 2022. **3, 7**
- [59] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CAAI International Conference on Artificial Intelligence*, pages 46–57. Springer, 2023. **2, 3**
- [60] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. **5**
- [61] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, pages 1–13, 2023. **1**
- [62] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. *IEEE CVPR*, pages 1–16, 2024.
- [63] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmm for visual scoring via discrete text-defined levels. In *ICML2024*, 2024. **3, 5, 6, 7**
- [64] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. **1, 2, 3**
- [65] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *CoRR*, abs/2303.14420, 2023. **2**
- [66] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageRe-ward: Learning and evaluating human preferences for text-to-image generation. *CoRR*, abs/2304.05977, 2023. **1, 2, 3, 7**
- [67] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. **3**
- [68] Jiquan Yuan, Fanyi Yang, Jihe Li, Xinyan Cao, Jinming Che, Jinlong Lin, and Xixin Cao. Pku-aigciqa-4k: A perceptual quality assessment database for both text-to-image and image-to-image ai-generated images. *arXiv preprint arXiv:2404.18409*, 2024. **2, 3**
- [69] Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał K Mantiuk, and Frédéric Dufaux. The relation between mos and pairwise comparisons and the importance of cross-content comparisons. *Electronic Imaging*, 30:1–6, 2018. **3**
- [70] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE CVPR*, pages 586–595, 2018. **3**
- [71] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023. **1, 3**
- [72] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. Q-bench-video: Benchmarking the video quality understanding of lmm. *arXiv preprint arXiv:2409.20063*, 2024. **2**
- [73] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al. Q-boost: On visual quality assessment ability of low-level multi-modality foundation models. In *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2024. **1, 6**
- [74] Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. A-bench: Are lmm masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*, 2024. **2, 3**
- [75] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *arXiv preprint arXiv:2402.07116*, 2024. **2**
- [76] Zicheng Zhang, Haoning Wu, Yingjie Zhou, Chunyi Li, Wei Sun, Chaofeng Chen, Xiongkuo Min, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. Lmm-pcqa: Assisting point cloud quality assessment with lmm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7783–7792, 2024. **2**
- [77] Zicheng Zhang, Yingjie Zhou, Chunyi Li, Baixuan Zhao, Xiaohong Liu, and Guangtao Zhai. Quality assessment in the era of large models: A survey. *arXiv preprint arXiv:2409.00031*, 2024. **2, 3**

Q-Eval-100K: Evaluating Visual Quality and Alignment Level for Text-to-Vision Content

Supplementary Material

7. Dataset Construction Details

In this section, we mainly talk about the details of prompts collection.

7.1. Prompts Collection

The prompt collection comprises two sources:

- **Internally Constructed Prompts**, which is based on internal capability requirement of Q-Eval-100K.
- **Open-Source Prompts**, which is based on other text-to-vision alignment evaluation datasets, including GenAIBench [34] and Docci [47]. GenAIBench features comprehensive prompt designs, while Docci provides longer prompts, making it suitable for evaluating long-prompt descriptions.

7.2. Internally Constructed Prompts

- **Manual Construction:** Data is manually created by searching for commonly used prompts and rewriting them to align with the distribution of specific capabilities.
- **GPT-4 Augmentation:** GPT-4 is used to expand the dataset for specific capabilities. This process involves leveraging a few manually constructed examples and applying a Chain-of-Thought (CoT) approach. GPT-4 generates prompts based on given definitions and examples, which are then filtered and refined manually.

Example GPT-4 Prompt Generation Instruction:

You are an expert at crafting text-to-image prompts. I need prompts for text-to-image models based on the following category labels. Each label is explained with a description, the text before the ';' is the label, and the text after the ';' provides details. Use your imagination and creativity to generate relevant English prompts. The prompt length should be between [len_min] and [len_max]. Avoid extra content, only output prompts.

7.3. Prompt Designing

As shown in Table 7, the prompt design focuses on three main aspects: **Entity Generation**, **Entity Attribute Generation**, and **Interaction Capability**. 1) Entity generation targets the primary entities (*people, objects, etc.*) to be generated. 2) Entity attribute generation emphasizes the attributes (*clothing, color, material, visual quality, etc.*) of the entities. 3) Interaction capability focuses on the interactions between the generated entities and other entities or the background, such as their spatial relationships and actions.

Table 7. Detailed Descriptions of Entity Generation, Entity Attribute Generation, and Interaction Ability

Category	Subcategory	Count	
Entity Generation	Simple Entity Generation	1439	
	Simple Human Generation	/	
	Simple Object Generation	/	
	Other Simple Entity Generation	/	
	Complex Entity Generation	1729	
	Character Information Generation	/	
	Text and Symbol Generation	/	
	Chart Generation	/	
	Entity Attribute Generation	Basic Entity Attributes	1656
		Entity Shape Generation	/
Entity Position Generation		/	
Entity Color Generation		/	
Entity State Generation		/	
Other Entity Attributes Generation		/	
Person and Animal Attributes Generation		1500	
Emotion Generation		/	
Action Generation		/	
Specific Age Person Generation		/	
Specific Gender Person Generation		/	
Other Person and Animal Attributes		/	
Portrait Generation		531	
Simple Portrait Generation		/	
Complex Portrait Generation		/	
Scene and Theme Generation		2450	
Theme Generation		/	
Scene Generation		/	
Style Generation		294	
Basic Visual Attributes Generation	321		
Image Sharpness Generation	/		
Exposure Generation	/		
Lighting Generation	/		
Contrast Generation	/		
Color Saturation Generation	/		
Noise Level Generation	/		
Composition Generation	/		
Color Balance Generation	/		
Depth of Field Generation	/		
Perspective Generation	/		
Camera Angle Generation	/		
Other Basic Visual Attributes Generation	/		
Interaction Ability	Interacting Multi-Entity Generation	1729	
	Sequential Relationship Multi-Entity Generation	/	
	Causal Relationship Multi-Entity Generation	/	
	Spatial Relationship Multi-Entity Generation	/	

8. Long Prompt Split

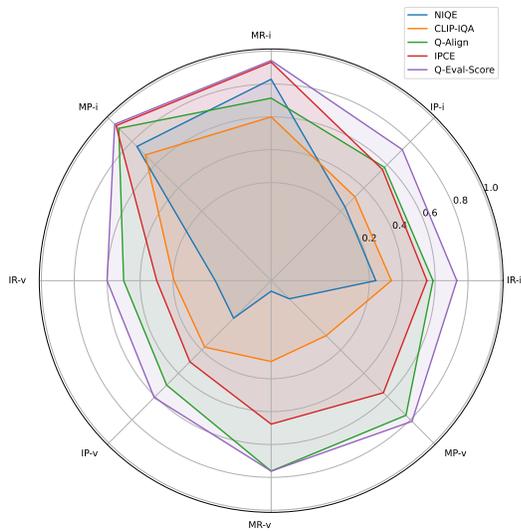
We use Qwen-VL-72B-Instruct [6] to help summarize the long prompt and split the long prompt into short sentences. Specifically, the prompt is designed as follows:

Summarize Prompt

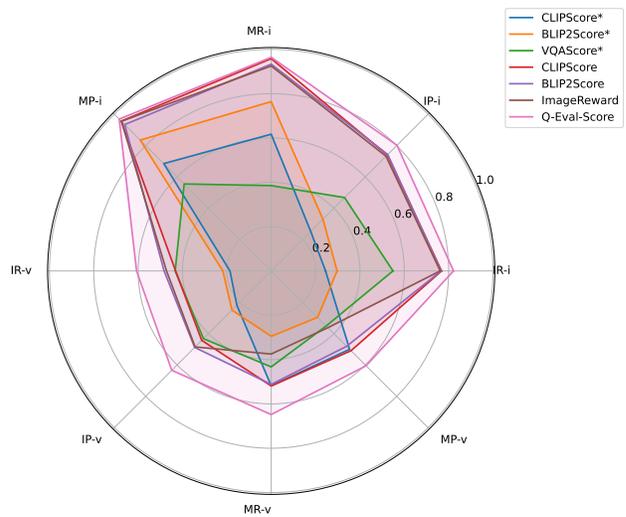
User: Please shorten the prompt to between 15 and 25 words, retaining the main information and ignoring details, specifically the characters, attributes, actions, and scenes. The prompt is as follows [Prompt].

Split Prompt

User: Split the prompt into three or fewer shorter prompts, with each short prompt describing one aspect of the original long prompt's subject and should be fewer than 15 words. The prompt is as follows [Prompt].



(a) Overall performance on **Visual Quality**.

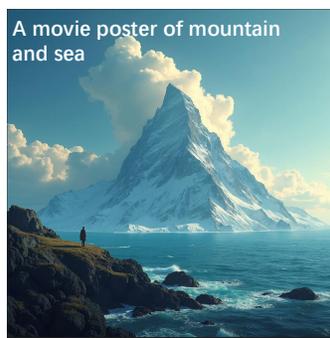


(b) Overall performance on **Alignment**.

Figure 7. Radar charts of the overall performance on the **Visual Quality** and **Alignment** aspects on **Q-Eval-100K**, where IR, IP, MR, MP indicate Instance-level SRCC, Instance-level PLCC, Model-level SRCC, Model-level PLCC, and -, v represents image and video respectively.



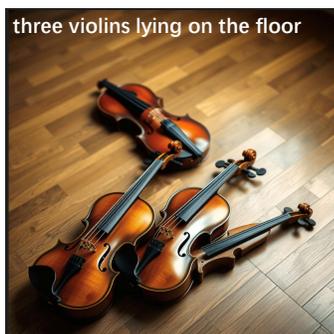
MOS	2.4
CLIPScore	4.5
BLIP2Score	4.6
Q-Eval-Score	2.6



MOS	4.1
CLIPScore	2.4
BLIP2Score	2.6
Q-Eval-Score	3.9



MOS	4.8
CLIPScore	2.5
BLIP2Score	2.9
Q-Eval-Score	4.6



MOS	3.5
CLIPScore	4.7
BLIP2Score	4.8
Q-Eval-Score	3.4



MOS	4.2
CLIPScore	2.2
BLIP2Score	2.1
Q-Eval-Score	4.0



MOS	4.4
CLIPScore	1.5
BLIP2Score	2.2
Q-Eval-Score	4.3

Figure 8. Visualization comparison results.

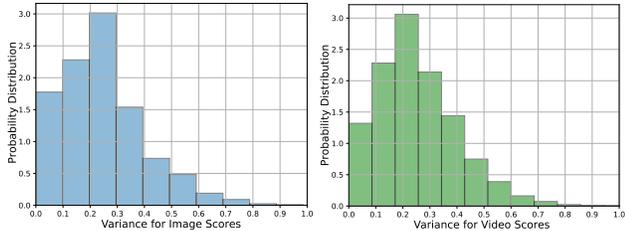


Figure 9. Variance probability distributions for images/videos of Q-Eval-100K respectively.

8.1. Subjective experiment details:

1) Each instance in the training and test sets is rated by at least 3 and 12 individuals on average. 2) We ensure raters’ diversity by employing raters from a wide age range (18-52) and selecting raters from various professional backgrounds. Each rater annotates a maximum of 30 instances at a time, followed by a mandatory 30-minute break. 3) Perfect-score cases are rated by 12 individuals first, then reviewed and adjusted by a group of 5 experts. 4) Given the scale of Q-Eval-100K (*the largest AIGC QA dataset with MOS at the cost of about 150,000 US dollars in total*), involving 15 raters per instance as suggested by ITU [1] would be impractical due to time and cost constraints. To preserve the dataset’s scale (*crucial for LMM training under scaling laws*), we reduce the number of raters and implement a ‘*Sample and Scrutinize*’ approach to maintain annotation quality. 5) The variance distribution of instance annotations is shown in Fig. 9, where most instances have a variance below 0.3.

9. Performance Details

9.1. Radar Charts of Overall Performance

To provide a comprehensive overview of the performance, we present the radar charts in Fig. 7. The key observations are as follows:

- **Visual Quality.** The proposed **Q-Eval-Score** outperforms all competitors, achieving the highest overall performance. The slight decline in video performance can be attributed to the 1fps frame sampling method, which reduces temporal information and affects accuracy. Despite this limitation, **Q-Eval-Score** leads the second-best competitor (Q-Align) by a notable margin of 10% in video instance-level SRCC.
- **Alignment.** **Q-Eval-Score** also excels in alignment evaluation, surpassing competitors by 6% in image instance-level SRCC and 12% in video instance-level SRCC. Furthermore, the significant performance gains seen in other trained models indicate that **Q-Eval-100K** serves as a valuable dataset for improving alignment evaluation methods.
- **Comparison Between Tasks.** The performance of **Q-Eval-Score** in visual quality evaluation is relatively lower than in alignment tasks, highlighting the greater complex-

ity of predicting visual quality. Alignment evaluation is more straightforward and objective, while visual quality involves nuanced and subjective perception, making it more challenging to assess.

Overall, the proposed **Q-Eval-Score** demonstrates remarkable capabilities in both visual quality and alignment evaluation. With performance exceeding 0.94 in image model-level metrics, it aligns closely with human judgment. These results underscore the robustness of **Q-Eval-Score** and its potential as a highly effective evaluation metric.

9.2. More Cross-validation Results

We further select 4 datasets for cross-validation (See Table 8). AGIQA^{Quality} [35] & T2VQA [29] are for visual quality, while AGIQA^{Align} [35] & TIFA160 [23] are for text alignment. The results show good generalization ability of Q-Eval-Score. (best in **bold**)

Table 8. Cross-validation (All pre-trained on Q-Eval-100K).

Dataset	AGIQA ^{Quality}	T2VQA	AGIQA ^{Align}	TIFA160
Method	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC	SRCC/PLCC
Q-Align	0.6581/0.6743	0.2539/0.2198	<i>Inapplicable</i>	<i>Inapplicable</i>
CLIPScore	<i>Inapplicable</i>	<i>Inapplicable</i>	0.5515/0.5627	0.5903/0.5952
BLIP2Score	<i>Inapplicable</i>	<i>Inapplicable</i>	0.6873/0.7085	0.7267/0.7468
Q-Eval-Score	0.7256/0.7248	0.4479/0.4498	0.7544/0.7432	0.7845/0.7954

9.3. Visualization Results

We provide additional comparison examples in Fig. 8 to offer a clearer understanding of the evaluation capabilities of different models. It is evident from these examples that both CLIPScore and BLIPScore struggle significantly in tasks such as recognizing text within images and accurately counting objects. These models often fail by assigning disproportionately high scores to results that do not align with the intended outputs, reflecting their limitations in fine-grained assessment. Furthermore, when dealing with complex scenarios involving long and detailed prompts, these models exhibit a consistent tendency to assign significantly lower alignment scores, likely due to their inability to effectively parse and match intricate contextual information. In contrast, **Q-Eval-Score** consistently demonstrates a much higher degree of accuracy and reliability in these challenging scenarios. These results further highlight the potential of **Q-Eval-Score** as a unified framework for evaluating text-to-vision generative models across diverse and demanding conditions.

10. Data Statement

Considering the large scale of the dataset and the complexity of the model, we are actively organizing and refining the content to ensure its quality and usability. We solemnly pledge to release the **Q-Eval-100K** dataset in carefully

planned batches, ensuring a comprehensive and systematic open-sourcing process that effectively supports community development. **Furthermore, we confirm that the dataset has successfully passed ethical review, affirming our commitment to responsible AI practices.** Alongside the dataset, we will also release **Q-Eval-Score** and provide continuous updates, ensuring the model remains aligned with the rapid advancements in generative AI.

11. Broader Impact and Limitations

11.1. Broader Impact

- **Empowering Generative AI Applications.** The development of comprehensive evaluation methods, such as Q-Eval-100K and Q-Eval-Score, directly supports these advancements by ensuring the quality and alignment of generated content, enabling its effective deployment.
- **Driving Standardization in Evaluation.** By introducing a unified framework for assessing visual quality and alignment, this work provides a benchmark for systematic evaluation. This standardization not only enhances the reliability of evaluations across diverse use cases but also fosters transparency in generative AI systems.
- **Facilitating Improvements in Generative Models.** The dataset and framework encourage the refinement of generative models by providing detailed feedback on visual quality and alignment. These insights guide iterative improvements, pushing the boundaries of what generative AI can achieve.

11.2. Limitations

- **Subjectivity in Visual Quality Evaluation.** While Q-Eval-Score aligns closely with human evaluations, the inherently subjective nature of visual quality perception may result in variability. Differences in individual preferences and cultural factors could affect the generalizability of the evaluation framework.
- **Dependency on Human Annotations.** The reliance on extensive human annotations for dataset creation introduces scalability issues and potential biases. Automating parts of this process without sacrificing quality remains an open challenge.