

# LMM4LMM: Benchmarking and Evaluating Large-multimodal Image Generation with LMMs (Supplementary Material)

Jiarui Wang<sup>1</sup>, Huiyu Duan<sup>1</sup>, Yu Zhao<sup>1</sup>, Juntong Wang<sup>1</sup>, Guangtao Zhai<sup>1</sup>, Xiongkuo Min<sup>1\*</sup>,  
<sup>1</sup>Shanghai Jiao Tong University, Shanghai, China

## 1. Overview

In this supplementary material, we provide additional details on the data collection, methodology, experiments, and results discussed in the main paper. In data collection, we detail the 20 distinct tasks in Section 2 and the overview of the 24 LMM-T2I models in Section 3. We then elaborate on the subjective experiments in Section 4, including the annotation dimension, criteria, interface and management. In addition, we provide an in-depth analysis of the EvalMi-50K database, including MOS distributions and model performance comparisons across the 20 tasks in Section 5. We outline the loss functions used in the training process for the LMM4LMM model in Section 6. Details on the evaluation criteria and algorithms are also included in Section 7. Finally, we provide more performance comparisons between our model and other metrics in Section 8.

## 2. T2I Task-specific Challenge Define

In this study, we systematically investigate the capabilities of text-to-image (T2I) generation models through a comprehensive evaluation framework. We focus on 20 distinct tasks that vary in complexity and require diverse compositional skills, as detailed in Table 1 with their corresponding subcategories, keywords, and example prompts. These tasks are carefully designed to assess different aspects of model performance, ranging from basic object rendering to complex spatial and attribute understanding, as shown in Figures 11-15. Below, we provide an overview of the main task categories and their associated challenges.

- **Single class:** evaluates a model’s ability to generate a single instance of a specified object class. The challenge lies in producing high-fidelity representations that maintain essential class-specific features without additional contextual constraints.
- **Two class:** evaluates a model’s capacity to simultaneously render two distinct object classes within a single image. This task introduces the challenge of maintaining object integrity while managing inter-object relationships. The complexity increases when considering potential occlusions, relative scaling, and basic spatial arrangements between the two objects.

- **Counting:** evaluates a model’s ability to generate a specific number of objects in a scene. The challenge includes numerical understanding and managing multiple instances without overlap or spatial issues, especially for larger numbers.
- **Colors:** evaluates a model’s proficiency in associating specific color attributes with generated objects. The challenge lies in accurately binding color properties to target objects while maintaining object integrity and distinguishing foreground objects from background elements.
- **Position:** evaluates a model’s capability to render two objects with specified positional relationships. The challenge encompasses not only object generation but also the accurate representation of specific spatial relationships (*e.g.*, above, below, left of, right of). This requires precise control over object arrangement while maintaining their identities.
- **Shapes:** evaluates a model’s ability to generate objects with specific geometric shapes (*e.g.*, spherical, rectangular, triangular, star) while preserving their recognizability. This tests the ability to abstract representations of real-world objects and express them in other shapes.
- **Texture:** evaluates a model’s capability to render objects with specific surface textures and material properties (*e.g.*, metallic, wooden, glass). The challenge lies in creating realistic textures that match the object’s properties and lighting conditions.
- **Scene:** evaluates a model’s ability to create complex scenes with multiple naturally composed elements in a specific environment (*e.g.*, beach, forest, kitchen). The challenge is to ensure all objects and backgrounds are contextually relevant and spatially consistent, evaluating the model’s holistic scene understanding.
- **Style:** evaluates a model’s proficiency in generating images in specific artistic styles (*e.g.*, watercolor, oil painting, cartoon). The challenge is to mimic the style’s visual characteristics while keeping objects and scenes recognizable, testing the model’s ability to apply abstract stylistic concepts consistently.
- **OCR (Optical Character Recognition):** evaluates a model’s capability to generate readable text within images, such as words or short sentences. The challenge is to make the text visually coherent with the image and machine-readable by OCR systems, testing the model’s

---

\*Corresponding Author

Table 1. Prompt categories with corresponding keywords and examples.

Category	Subcategory / Keywords	Prompt examples
Single Class	person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, ...	A photo of a bench
Two Class	bench & sports, sheep & dog, cow & elephant, knife & spoon, chair & couch, ...	A photo of a bench and a sports ball
Counting	zero, one, two, three, four, five, six, seven, eight, nine, ten	A photo of three computer keyboards
Colors	red, orange, yellow, green, blue, purple, pink, brown, black, white	A photo of a black donut
Position	left of, right of, above, below	A photo of a bottle right of a train
Shapes	circle, cylinder, sphere, star, triangle, rectangle, irregular, oval, linear, cone	A photo of a circle skateboard
Texture	glass, cement, stone, rubber, fabric, ceramics, leather, metallic, wooden, plastic	A photo of a fabric model bicycle
Scene	kitchen, living room, street, swimming pool, playground, waterfall, forest	A photo of in the forest
Style	cartoon, realistic, oil painting, vintage, watercolor, line drawing	A vintage image of a tv remote
OCR	“HELLO”, “STOP”, “SUCCESSFUL”, “Have a nice day”, “Enjoy life”, “Keep going”, ...	A photo of phrase “Believe in yourself”
HOI	hold a stop sign, operate an oven, peel an apple, lie on a bench, carry a book, ...	A photo of people boarding a car
Human	human, cloth, cloth-color, hair, hair-color	A man in a blue shirt smiles warmly, his curly black hair framing his face
Emotion	happy, sadness, love, fear, surprise, anger, worry, neutrality	A dog is smiling with happy emotion. He finds a lot of delicious food
Linguistic Structure	without, no, not	The garden has no flowers blooming. It is late in the winter
View	close-up, ground view, aerial view, overhead view, first-person view, wide-angle view, ...	An overhead view of a pickup truck with boxes in its flatbed
World Knowledge	Great Wall, Great Pyramid, Ha Long Bay, Machu Picchu, Eiffel Tower, Grand Canyon, ...	boats in Ha Long Bay
Face	hair, mouth, emotion, eyes, necklace, cheeks, nose, skin	A face image with medium length hair, wearing necklace
Imagination	—	A panda is flying in the sky
Time & Light	time: sunset, early morning, night, midnight, midday, noon, dawn, ... light: fiery orange, golden, moonlight, silvery, misty, bright, crimson, ...	As the sun sets, fiery orange light streaks across the sky, casting a warm glow over the city skyline and the distant hills
Complex	Counting + Color + Shapes + Scene, Style + Color + Position, Human + Emotion, ...	A photo of four blue birds playing on a circle playground

understanding of typography and text integration.

- **HOI (Human-Object Interaction):** evaluates a model’s ability to generate realistic interactions between humans and objects, ensuring the actions are physically plausible. The challenge is to create recognizable humans and objects while maintaining natural spatial and logical relationships.
- **Human:** evaluates a model’s ability to generate human figures with specific occupational attire, unique accessories, and hairstyles. The challenge lies in creating realistic and coherent human representations while maintaining consistency across these attributes.
- **Emotion:** evaluates a model’s ability to convey specific emotions or moods, either through human facial expressions (*e.g.*, happiness, sadness) or through the overall atmosphere of a scene (*e.g.*, serene, love). This evaluates the model’s understanding of emotional cues and its ability to translate abstract emotions into visuals.
- **Linguistic Structure:** evaluates a model’s ability to interpret and render linguistic structures involving negation (*e.g.*, “without,” “no”). The challenge is to generate images that accurately reflect the absence of specified objects or features (*e.g.*, a “classroom without people”) while maintaining scene integrity. This tests the model’s comprehension of negative constructs.
- **View:** evaluates a model’s ability to generate images from

specific viewpoints (*e.g.*, first-person, third-person, side view). The challenge is to maintain correct spatial orientation, scale, and proportion across perspectives, testing the model’s understanding of spatial geometry.

- **World Knowledge:** evaluates a model’s knowledge of real-world landmarks, historical sites (*e.g.*, the Great Wall, Eiffel Tower, Great Pyramid), and the physical appearances of famous individuals (*e.g.*, Albert Einstein). The challenge lies in creating content that accurately aligns with people’s perceptions of famous landmarks and the physical appearances of well-known individuals.
- **Face:** evaluates a model’s ability to generate human faces with specific features (*e.g.*, face shape, nose structure, hairstyle). The challenge is to create realistic and diverse facial representations while maintaining feature consistency, and testing the model’s understanding of facial anatomy.
- **Imagination:** evaluates a model’s ability to generate imaginative scenes that combine elements from different categories or depict impossible scenarios in the real world (*e.g.*, a “cat wearing a chef’s hat cooking in a kitchen”). The challenge is to balance creativity with visual plausibility, evaluating the model’s capacity for creative thinking and novel concept synthesis.
- **Time & Light:** evaluates a model’s ability to generate images that accurately depict different times of day (*e.g.*,

morning, evening) and lighting conditions (e.g., sunlight, dim light). The challenge is to adjust brightness, color temperature, shadows, and reflections appropriately and test the model’s understanding of time-based lighting dynamics and its ability to visually represent them.

- **Complex:** is designed by combining simpler task components, such as color recognition, object counting, and shape identification, into more intricate and multifaceted challenges. These tasks require models to integrate and execute multiple simple tasks simultaneously within a single image. Below are some combined forms of complex tasks along with corresponding examples:
- (1) **Counting + Color + Shapes + Scene:** A photo of [number] [color] [class] [action] in a [shape] [scene]. **Example:** *A photo of two white dogs swimming in a triangle-shaped swimming pool.*
  - (2) **Counting + Color + Shapes + Texture:** A photo of [number] [color] [texture] [shape] [class]. **Example:** *A photo of two brown wooden rectangular books.*
  - (3) **HOI + Color + Shape + Texture:** A photo of [human action] a [color] [texture] [shape] [object]. **Example:** *A photo of people opening a yellow wooden triangle box.*
  - (4) **Style + Color + Position:** A [style] image of a [color1] [class1] [position] a [color2] [class2]. **Example:** *A cartoon image of a yellow dog to the left of a white cat.*
  - (5) **Style + OCR + Color:** A [style] image of [color] text “[content]”. **Example:** *An oil painting of red text “CONGRATULATIONS”.*
  - (6) **OCR + Color + Single Class:** A photo of [color1] text “[content]” on a [color2] [class]. **Example:** *A photo of green text “Happy Birthday” on a pink cake.*
  - (7) **Counting + Shapes + Two Classes:** A photo of [number1] [shape1] [class1] and [number2] [shape2] [class2]. **Example:** *A photo of six spherical balls and three rectangular cups.*
  - (8) **Counting + Color + Two Classes:** A photo of [number1] [color1] [class1] and [number2] [color2] [class2]. **Example:** *A photo of six red books and four blue pens.*
  - (9) **View + World Knowledge:** A [view] of [famous landmark]. **Example:** *An aerial view of the Great Wall.*
  - (10) **Human + Emotion:** A [human description] [action] with [emotion]. **Example:** *A girl in a white blouse and navy skirt, wearing a red ribbon tie, smiles with excitement as she receives a trophy during a school award ceremony. Her long brown hair shines as she turns to the audience.*

### 3. Detailed Information of T2I Models

**Stable Diffusion v2.1** [34] is a model designed for generating and modifying images based on text prompts. It is a Latent Diffusion Model that employs a fixed, pretrained

text encoder (OpenCLIP-ViT/H [33]). It is conditioned on the penultimate text embeddings of a CLIP ViT-H/14 [33] text encoder.

**i-Code-V3** [39] is a composable diffusion model capable of generating language, image, video, and audio from any input combination. It reuses Stable Diffusion 1.5’s structure and weights, leveraging large-scale datasets like LAION-400M to achieve high-quality multi-modal generation with strong cross-modal coherence.

**Stable Diffusion XL (SDXL)** [32] massively increases the UNet backbone size from Stable Diffusion v2 [34] and incorporates two text encoders. There is a second refinement model, which we do not use as it does not affect the composition of the image.

**DALLE3** [3], developed by OpenAI, enhances spatial reasoning and improves the handling of complex prompts by leveraging advanced transformer architectures and refined training datasets, enabling the generation of highly detailed and contextually accurate images.

**LLMGA** [48] enhances multimodal large language models (LLMs) by generating detailed text prompts for Stable Diffusion (SD) [34], improving contextual understanding and reducing noise in generation. It leverages a diverse dataset for prompt refinement, image editing, and inpainting, enabling more precise and flexible image synthesis.

**Kandinsky-3** [1] is a hybrid model combining diffusion and transformer architectures, which emphasizes artistic and abstract image generation. It is particularly effective for creating visually striking and imaginative compositions.

**LWM** [28] is a multimodal autoregressive model trained on extensive video and language data. Using RingAttention, it efficiently handles long-sequence training, expanding context size up to 1M tokens, enabling strong language, image, and video understanding and generation.

**Playground** [24] is designed for high-resolution and photorealistic outputs, which incorporates advanced noise scheduling and fine-tuning techniques. It is optimized for generating detailed and visually appealing images with minimal artifacts.

**LaVi-Bridge** [57] is designed for text-to-image diffusion models and serves as a bridge, which enables the integration of diverse pre-trained language models and generative vision models for text-to-image generation. By leveraging LoRA and adapters, it offers a flexible and plug-and-play approach without requiring modifications to the original weights of the language and vision models.

**ELLA** [18] is a method that enhances current text-to-image diffusion models with state-of-the-art large language models (LLMs) without requiring the training of LLMs or U-Net. We design a lightweight and adaptive Timestep-Aware Semantic Connector (TSC) to effectively condition the image generation process, ensuring comprehensive prompt understanding from the LLM. With ELLA, the diffusion

Table 2. An overview and URLs of the adopted 24 text-to-image generation models.

Models	Type	Date	Resolution	URL
SD_v2-1 [34]	Diff.	2022.12	768×768	<a href="https://huggingface.co/stabilityai/stable-diffusion-2-1">https://huggingface.co/stabilityai/stable-diffusion-2-1</a>
i-Code-V3 [39]	Diff.	2023.05	256×256	<a href="https://github.com/microsoft/i-Code">https://github.com/microsoft/i-Code</a>
SDXL_base-1 [32]	Diff.	2023.07	1024×1024	<a href="https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0">https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0</a>
DALLE3 [3]	Diff.	2023.09	1024×1024	<a href="https://openai.com/index/dall-e-3">https://openai.com/index/dall-e-3</a>
LLMGA [48]	Diff.	2023.11	1024×1024	<a href="https://github.com/dvlab-research/LLMGA">https://github.com/dvlab-research/LLMGA</a>
Kandinsky-3 [1]	Diff.	2023.12	1024×1024	<a href="https://github.com/ai-forever/Kandinsky-3">https://github.com/ai-forever/Kandinsky-3</a>
LWM [28]	AR	2024.01	256×256	<a href="https://github.com/LargeWorldModel/LWM">https://github.com/LargeWorldModel/LWM</a>
Playground [24]	Diff.	2024.02	1024×1024	<a href="https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic">https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic</a>
LaVi-Bridge [57]	Diff.	2024.03	512×512	<a href="https://github.com/ShihaoZhaoZSH/LaVi-Bridge">https://github.com/ShihaoZhaoZSH/LaVi-Bridge</a>
ELLA [18]	Diff.	2024.03	512×512	<a href="https://github.com/TencentQGYLab/ELLA">https://github.com/TencentQGYLab/ELLA</a>
Seed-xi [11]	Diff.	2024.04	1024×1024	<a href="https://github.com/AILab-CVC/SEED-X">https://github.com/AILab-CVC/SEED-X</a>
PixArt-sigma [5]	Diff.	2024.04	1024×1024	<a href="https://github.com/PixArt-alpha/PixArt-sigma">https://github.com/PixArt-alpha/PixArt-sigma</a>
LlamaGen [36]	AR	2024.06	256×256	<a href="https://github.com/FoundationVision/LlamaGen">https://github.com/FoundationVision/LlamaGen</a>
Kolors [40]	Diff.	2024.07	1024×1024	<a href="https://github.com/Kwai-Kolors/Kolors">https://github.com/Kwai-Kolors/Kolors</a>
Flux.schnell [22]	Diff.	2024.08	1024×1024	<a href="https://huggingface.co/black-forest-labs/FLUX.1-schnell">https://huggingface.co/black-forest-labs/FLUX.1-schnell</a>
Omnigen [49]	Diff.	2024.09	1024×1024	<a href="https://github.com/VectorSpaceLab/OmniGen">https://github.com/VectorSpaceLab/OmniGen</a>
EMU3 [42]	AR	2024.09	720×720	<a href="https://github.com/baaivision/Emu">https://github.com/baaivision/Emu</a>
Vila-u [46]	AR	2024.09	256×256	<a href="https://github.com/mit-han-lab/vila-u">https://github.com/mit-han-lab/vila-u</a>
SD3.5_large [10]	Diff.	2024.10	1024×1024	<a href="https://huggingface.co/stabilityai/stable-diffusion-3.5-large">https://huggingface.co/stabilityai/stable-diffusion-3.5-large</a>
Show-o [50]	AR+Diff.	2024.10	512×512	<a href="https://github.com/showlab/Show-o">https://github.com/showlab/Show-o</a>
Janus [43]	AR	2024.10	384×384	<a href="https://github.com/deepseek-ai/Janus">https://github.com/deepseek-ai/Janus</a>
Hart [38]	AR	2024.10	1024×1024	<a href="https://github.com/mit-han-lab/hart">https://github.com/mit-han-lab/hart</a>
NOVA [9]	AR	2024.12	512×512	<a href="https://github.com/baaivision/NOVA">https://github.com/baaivision/NOVA</a>
Infinity [14]	AR	2024.12	1024×1024	<a href="https://github.com/FoundationVision/Infinity">https://github.com/FoundationVision/Infinity</a>

model can generate high-fidelity and accurate images based on long, information-dense prompts.

**Seed-xi** [11] is a unified and versatile foundation model that can serve as a multimodal AI assistant in real-world applications. Through different instruction tuning, it can respond to various user needs by unifying multi-granularity comprehension and generation.

**PixArt-sigma** [5] is a Diffusion Transformer model (DiT) capable of directly generating images at 4K resolution. Representing a significant advancement over its predecessor, PixArt-alpha [4], it offers markedly higher image fidelity and improved alignment with text prompts. A key feature of PixArt-sigma [5] is its training efficiency.

**LlamaGen** [36] applies the next-token prediction paradigm of large language models to image generation. By refining image tokenizers and training datasets, it surpasses diffusion models in class-conditional generation and maintains competitive text alignment in text-to-image synthesis.

**Kolors** [40] is a large-scale latent diffusion model developed by the Kuaishou Kolors team for text-to-image generation. Trained on billions of text-image pairs, it outperforms both open-source and closed-source models in visual quality, complex semantic accuracy, and text rendering. Supporting both Chinese and English inputs, it excels at generating high-fidelity images while demonstrating strong performance in understanding Chinese-specific content.

**Flux.schnell** [22] is a 12 billion parameter rectified flow transformer capable of generating images from text descriptions. Trained using latent adversarial diffusion distillation,

it can generate high-quality images in only 1 to 4 steps. The model is very responsive and suitable for personal development

**OmniGen** [49] is a unified image generation model capable of producing a wide range of images from multi-modal prompts. It is designed to be simple, flexible, and easy to use. As a new diffusion model for unified image generation, it not only excels in text-to-image generation but also inherently supports various downstream tasks, such as image editing, subject-driven generation, and visual conditional generation.

**EMU3** [42] is a multimodal model that leverages next-token prediction as its sole training paradigm. By tokenizing images, text, and videos into a unified discrete space, it enables a single Transformer to be trained from scratch on diverse multimodal sequences. It streamlines the multimodal learning process, enhancing both efficiency and versatility in handling complex multimodal interactions.

**Vila-u** [46] is a unified foundation model for video, image, and language understanding and generation. Unlike traditional VLMs with separate modules, it employs a single autoregressive framework, simplifying architecture while achieving near state-of-the-art performance in both comprehension and generation.

**Stable Diffusion 3.5 Large** [10] is a Multimodal Diffusion Transformer (MMDiT) text-to-image model that features improved performance in image quality, typography, complex prompt understanding, and resource-efficiency. It uses three fixed, pretrained text encoders, and with QK-



normalization to improve training stability.

**Show-o** [50] processes text tokens autoregressively with causal attention while handling image tokens using (discrete) denoising diffusion modeling via full attention. It then generates the desired output. Specifically, it is capable of performing image captioning, visual question answering, text-to-image generation, text-guided inpainting/extrapolation, and mixed-modality generation.

**Janus** [43] is a novel autoregressive framework that unifies multimodal understanding and generation. It addresses the limitations of previous approaches by decoupling visual encoding into separate pathways while still utilizing a single, unified transformer architecture for processing.

**HART** [38] introduces a hybrid tokenizer that enhances autoregressive (AR) models by improving image reconstruction quality and reducing training costs for high-resolution (1024px) image generation. It achieves this by decomposing the continuous latents from the autoencoder into two components: discrete tokens that capture the overall structure and continuous tokens that retain fine-grained residual details.

**NOVA** [9] is a model that enables autoregressive image/video generation with high efficiency. It reformulates the video generation problem as non-quantized autoregressive modeling of temporal frame-by-frame prediction and spatial set-by-set prediction. It generalizes well and enables diverse zero-shot generation abilities in one unified model.

**Infinity** [14] is a bitwise visual autoregressive model that adopts a novel token prediction framework with an infinite-vocabulary tokenizer and bitwise self-correction. By scaling the tokenizer vocabulary and transformer size concurrently, it enhances the model’s capacity for high-resolution image generation while maintaining fine-grained visual fidelity.

## 4. More Details of Subjective Experiment

### 4.1. Annotation Dimension and Criteria

To comprehensively assess the performance of AI-generated images (AIGIs), we propose a dual-dimensional evaluation framework that examines both perceptual quality and text-to-image (T2I) correspondence. This approach enables a thorough analysis of different aspects of image generation, providing a holistic understanding of a model’s capabilities and limitations.

- **Perceptual quality** evaluates the visual characteristics and aesthetic appeal of generated images. This dimension focuses on multiple aspects of image quality, including **visual clarity** (the sharpness and resolution of image details), **naturalness** (the degree to which the image appears realistic and free from artifacts), **aesthetic appeal** (the composition, color harmony, and overall visual attractiveness), **structural coherence** (the logical con-

sistency of spatial relationships and object proportions), and **authenticity** (whether the generated image is realistic). High-scoring images are characterized by exceptional clarity, vivid and well-balanced colors, and meticulous attention to detail, offering an immersive and visually striking experience. In contrast, low scores reflect images with blurriness, unnatural color tones, faded visuals, and a lack of clarity or detail. This dimension captures the foundational visual attributes that make an image aesthetically pleasing or distracting. For detailed criteria, refer to Figure 5.

- **Text-image correspondence** assesses the semantic alignment between the generated image and the input text prompt, including **content accuracy** (the presence and correct representation of described objects and elements), **contextual relevance** (the appropriate depiction of scenes and relationships between objects), **attribute fidelity** (the accurate representation of specific characteristics mentioned in the prompt), and **semantic consistency** (the logical coherence between visual elements and textual descriptions). Images with high scores perfectly match the descriptions in the prompt, accurately reflecting all elements with high fidelity. These images effectively translate textual information into visual content without mismatches. In contrast, images with lower scores exhibit inconsistencies, missing elements, or mismatched content. For detailed criteria, refer to Figure 6.

### 4.2. Significance of the Two Dimensions

The dual-dimensional evaluation framework, which combines perception quality and T2I correspondence, is essential for addressing the inherent trade-offs and complementary aspects of AIGIs. While perception quality emphasizes the visual characteristics that contribute to an image’s appeal and realism, T2I correspondence ensures that the generated content remains semantically faithful to the original textual description. Together, these dimensions provide a comprehensive assessment of both the aesthetic and functional aspects of image generation. As illustrated in Figure 1, a high perception quality score alone does not guarantee semantic accuracy. For example, an image may exhibit exceptional visual quality, characterized by high resolution, vibrant colors, and meticulous detail, yet fail to accurately represent the specific objects, relationships, or attributes described in the text prompt. Conversely, an image may perfectly align with the textual description in terms of content and context but suffer from poor visual quality, such as low resolution, unnatural textures, or inconsistent lighting, which detracts from its overall appeal and usability. The integration of both dimensions ensures that generated images achieve a balance between visual excellence and semantic fidelity. This holistic approach not only enhances the evaluation of generative models but also aligns

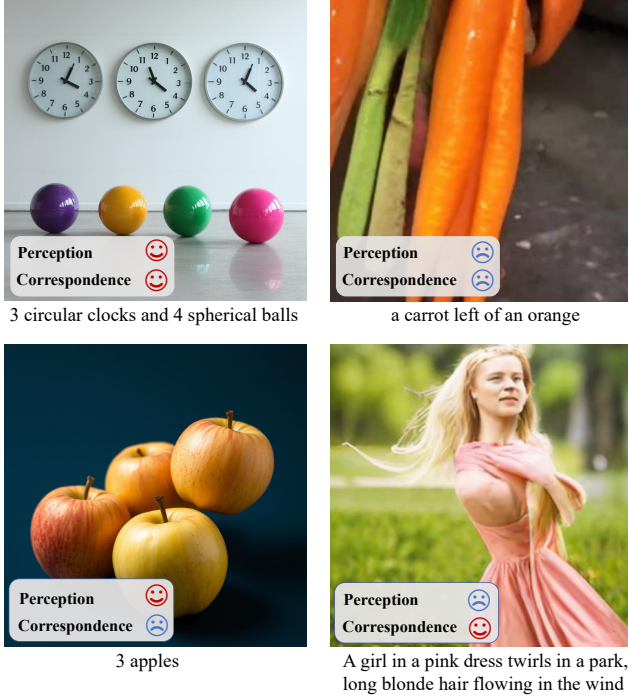


Figure 1. Illustration of the evaluation dimensions: perceptual quality and text-image correspondence, attached with examples with different subjective qualities.

with real-world applications where both image quality and content accuracy are critical. By considering both dimensions, the framework provides a more nuanced understanding of a model’s strengths and weaknesses, facilitating targeted improvements in image generation systems.

#### 4.3. Annotation Interface


To ensure a comprehensive and efficient image quality evaluation, we design two custom annotation interfaces tailored for different assessment tasks: simple task annotation and complex task annotation. The simple task annotation interface, shown in Figure 2, is a manual evaluation platform developed using the Python tkinter package, designed to facilitate MOS assessments. The experiment involves evaluating images based on two independent dimensions and answering a binary question related to a specific task-specific challenge. There are 20 task-specific challenges, including categories such as human, shape, scene, color, etc. Each trial presents three images that correspond to the same prompt. These images are randomly selected from 24 different models. Importantly, participants are instructed to assign absolute scores to each image on the two predefined dimensions, rather than making relative comparisons between the images. For each image, participants provide: (1) Two separate scores representing the two evaluation dimensions. (2) A binary response (yes/no) to indicate whether the image meets the specified challenge criterion. Meanwhile, the

complex task annotation interface, is illustrated in Figure 3. The complex tasks are composed of multiple subtasks such as Number, Color, Shape, and Scene. Each subtask is evaluated independently with a yes/no response. The complex task is considered correct only if all its sub-tasks are correct. If any sub-task is incorrect, the entire complex task is marked as incorrect. To ensure uniformity and minimize resolution-related biases in image quality evaluation, all images displayed in this interface are cropped to a spatial resolution of 1024×1024 pixels. Navigation options, such as “Previous” and “Next” streamline the workflow, enabling efficient annotation.

#### 4.4. Annotation Management

To ensure ethical compliance and the quality of annotations, we implement a comprehensive process for the EvalMi-50K dataset. All participants are fully informed about the experiment’s purpose, tasks, and ethical considerations. Each participant sign an informed consent agreement, granting permission for their subjective ratings to be used exclusively for noncommercial research purposes. The dataset, comprising 50,400 AIGIs alongside their corresponding prompts, has been publicly released under the CC BY 4.0 license, ensuring accessibility while adhering to ethical guidelines. We ensure the exclusion of all inappropriate or NSFW content (textual or visual) through a rigorous manual review process during the image generation phase. This step ensures that the dataset remains suitable for academic and research use. The annotation process is structured into two primary components: Mean Opinion Score (MOS) annotation and task-specific question-answering (QA) annotation. Each component is designed to evaluate images across 20 task-specific challenges, including color, position, shapes, view, and *etc.* The MOS annotation task involves 16 participants to rate each image on a 0-5 Likert scale, assessing both perception quality and T2I correspondence. The question-answering annotation task is similarly conducted with 16 participants, ensuring consistency in the evaluation process. In this task, participants are presented with a series of yes/no questions across the 20 task-specific challenges. To determine the final answer for each question, a majority voting mechanism is employed. This approach ensures that the final decision reflects the collective judgment of the participants, minimizing the impact of individual biases or errors.


Prior to engaging in the annotation tasks, all participants undergoes a rigorous training process. As illustrated in Figures 5-6, they are provided with detailed instructions and multiple standardized examples. To ensure a high level of understanding and consistency, a pre-test is conducted to evaluate participants’ comprehension of the criteria and their alignment with the standard examples. Participants who do not meet the required accuracy threshold are ex-



Perception  
0.0  
0 1 2 3 4 5

Correspondence  
0.0  
0 1 2 3 4 5


Human -- black suit -- short black hair  
☐ Yes ☐ No



Perception  
0.0  
0 1 2 3 4 5

Correspondence  
0.0  
0 1 2 3 4 5

Human -- black suit -- short black hair  
☐ Yes ☐ No



Perception  
0.0  
0 1 2 3 4 5


Correspondence  
0.0  
0 1 2 3 4 5

Human -- black suit -- short black hair  
☐ Yes ☐ No

A man stands confidently in a bustling city street, his posture relaxed yet assertive. He wears a sharp black suit that fits perfectly, and his short brown hair is neatly styled, giving him a polished, professional look. The noise and movement of the city seem to fade around him, as if he is a figure of calm amidst the chaos.

Previous
Next

Figure 2. An example of the simple task annotation interface for human evaluation. The subjects are instructed to rate two dimensions of AI-generated images, *i.e.*, perception and text-image correspondence, and provide a binary (yes/no) response for a task-specific challenge. Each trial presents three images generated from 24 models for the same prompt, with absolute scoring applied independently to each image.



Perception  
0.0  
0 1 2 3 4 5


Correspondence  
0.0  
0 1 2 3 4 5

Number -- 2 ☐ Yes ☐ No

Color -- white ☐ Yes ☐ No

Shape -- triangle ☐ Yes ☐ No

Scene -- swimming pool ☐ Yes ☐ No



Perception  
0.0  
0 1 2 3 4 5


Correspondence  
0.0  
0 1 2 3 4 5

Number -- 2 ☐ Yes ☐ No

Color -- white ☐ Yes ☐ No

Shape -- triangle ☐ Yes ☐ No

Scene -- swimming pool ☐ Yes ☐ No



Perception  
0.0  
0 1 2 3 4 5

Correspondence  
0.0  
0 1 2 3 4 5

Number -- 2 ☐ Yes ☐ No

Color -- white ☐ Yes ☐ No

Shape -- triangle ☐ Yes ☐ No

Scene -- swimming pool ☐ Yes ☐ No

A photo of two white dog swimming in a triangle swimming pool

Previous
Next

Figure 3. An example of the complex task annotation interface, which extends the simple task evaluation by incorporating multiple sub-tasks (*e.g.*, Number, Color, Shape, and Scene). The subjects are instructed to rate two dimensions of AI-generated images, *i.e.*, perception and text-image correspondence, based on the given image and its prompt. Each sub-task is judged independently with a yes/no response. The complex task is considered correct only if all sub-tasks are correctly identified; if any sub-task is incorrect, the entire complex task is marked as incorrect.



cluded from further participation, ensuring that only well-prepared individuals contribute to the final dataset. During the experiment, all evaluations are conducted in a controlled laboratory environment under normal indoor lighting conditions. Participants are seated at a comfortable viewing distance of approximately 60 cm from the screen to minimize visual strain and ensure consistent evaluation conditions. While individual preferences may naturally vary, the use of detailed explanations and standardized annotation criteria ensure a high degree of agreement among participants. This consensus is particularly evident in question-answering annotations, where majority voting effectively captures group preferences. This rigorous and ethically sound annotation management strategy establishes EvalMi-50K as a robust and reliable resource for advancing research in image quality assessment.

## 5. More Analysis of EvalMi-50K Database

### 5.1. MOS Distribution across 20 Challenges

As mentioned in the main text, we process and compute the valid subjective evaluation results, obtaining a total of 100,800 Mean Opinion Scores (MOSs) across two dimensions, along with QA accuracy. To better illustrate the generative capabilities of current T2I models in different prompt challenges, we categorize the computed MOSs data into 20 task categories and used the categorized data to plot histograms and kernel density curves (KDC) graphs, as shown in Figure 4. We can observe that the 24 T2I models we tested exhibit relatively poor text-image alignment in prompt challenges related to position, OCR, linguistic structures, and complexity, with MOSs primarily clustering around 30. In contrast, their performance in other prompt challenges is relatively better. The overall perception MOSs does not show significant differences across different prompt challenges, with scores generally concentrated at a higher level. However, models perform slightly worse in OCR, HOI, and Face-related prompt challenges, where lower MOSs appear more frequently compared to other prompt challenges.

### 5.2. T2I Model Performance across 20 Challenges

Tables 3-5 provide detailed performance comparisons of the 24 T2I models across 20 task-specific challenges on three types of human annotations: perception MOS, T2I correspondence MOS, and question-answering accuracy. For perception quality, as demonstrated in Table 3 and Figure 7-8, models like Playground [24] stand out with the highest MOS and perform particularly well in categories such as “Colors,” “Shapes,” and “Scene”. These models excel in generating images that are visually appealing, realistic, and aesthetically pleasing. For T2I correspondence, as demonstrated in Table 4 and Figure 9-10, SD3.5\_large [10] leads

the way, demonstrating strong alignment between the generated images and the textual descriptions, but has a relatively lower performance in perception quality. Conversely, models like Kolrs [40] excel in perception quality, delivering high MOS scores, but can not perform as well in terms of T2I correspondence. The contrasting trends in performance between perception quality and T2I correspondence emphasize the importance of evaluating both dimensions independently. While perception quality focuses on the visual aspects of the generated images, T2I correspondence measures how well the image aligns with the content described in the text prompt. This dual evaluation ensures a more comprehensive understanding of a model’s abilities, where one dimension evaluates aesthetic quality, and the other checks the accuracy of the image-text alignment. In terms of task-specific accuracy, as demonstrated in Table 5, the ranking of models largely mirrors the performance in T2I correspondence. Since task-specific accuracy is inherently tied to T2I correspondence, models that excel in faithfully translating text into images also tend to perform well in answering specific questions related to those images. While task-specific accuracy provides binary (0/1) assessments based on task-specific queries, MOS offers continuous scoring that enables a more granular evaluation of the text-image correspondence, providing deeper insights into how accurately a model generates images in relation to the given prompt, beyond a simple binary judgment.

## 6. Details of Loss Function

The training process for LMM4LMM is divided into two progressive stages, each utilizing a specific loss function to target distinct objectives: language loss for instruction tuning, aligning visual and language features to give visual question answers across the 20 task-specific challenges, L1 loss for quality regression fine-tuning to generate accurate perception and correspondence scores.

**(1) Instruction tuning with language loss.** In the first stage, we train the projector to align visual and language features using the standard language loss. This involves ensuring that the visual tokens extracted from the vision encoder correspond effectively to the language representations from the LLM. The language loss, calculated using a cross-entropy function, measures the model’s ability to predict the correct token given the prior context:

$$\mathcal{L}_{\text{language}} = -\frac{1}{N} \sum_{i=1}^N \log P(y_{\text{label}} | y_{\text{pred}}) \quad (1)$$

where  $P(y_{\text{label}} | y_{\text{pred}})$  represents the probability assigned to the correct token  $y_{\text{label}}$  by the model,  $y_{\text{pred}}$  is the predicted token, and  $N$  is the total number of tokens. By minimizing

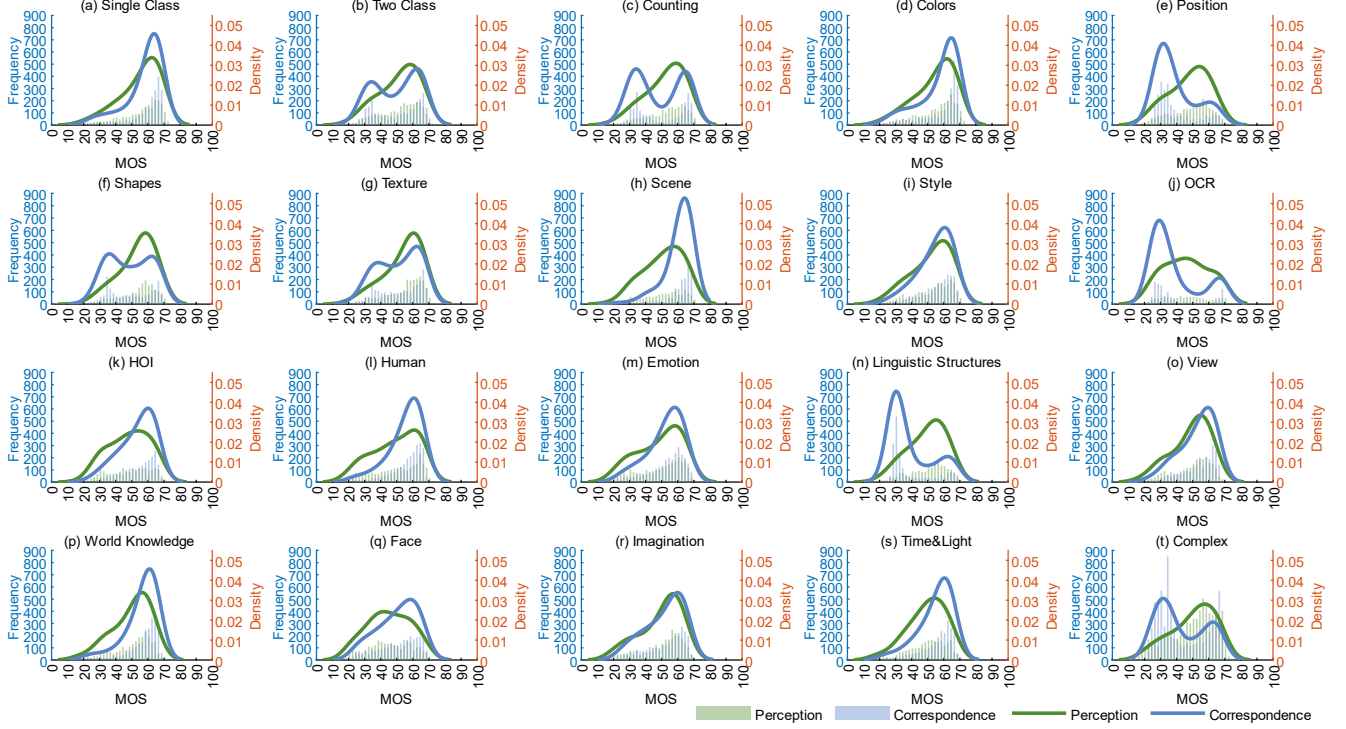


Figure 4. **Mean Opinion Score (MOS) distribution histograms and kernel density curves** of EvalMi-50K dataset. It includes two dimensions: Perception MOS and Correspondence MOS. Each dimension contains a total of 50,400 MOS values.

this loss, the model learns to generate coherent textual descriptions of image content, laying the foundation for subsequent stages.

**(2) Refining quality scoring with L1 loss.** Once the model can produce coherent descriptions of image content, the focus shifts to fine-tuning the quality regression module to output stable and precise numerical quality scores. The quality regression module takes the aligned visual tokens as input and predicts a quality score that reflects the overall image quality. Using the EvalMi-50K, which contains human-annotated MOS for each image, the model is trained to align its predictions with human ratings. The training objective minimizes the difference between the predicted quality score  $Q_{predict}$  and the ground-truth MOS  $Q_{label}$  using the L1 loss function:

$$\mathcal{L}_{MOS} = \frac{1}{N} \sum_{i=1}^N |Q_{predict}(i) - Q_{label}(i)| \quad (2)$$

where  $Q_{predict}(i)$  is the score predicted by the regressor  $i$  and  $Q_{label}(i)$  is the corresponding ground-truth MOS derived from subjective experiments, and  $N$  is the number of images in the batch. This loss function ensures that the predicted scores remain consistent with human evaluations, enabling the model to accurately assess the quality of AI-generated images in numerical form.

## 7. Implementation Details

### 7.1. Detailed Information of Evaluation Criteria

We adopt the widely used metrics in IQA literature [37, 44, 56]: Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), and Kendall’s Rank Correlation Coefficient (KRCC) as our evaluation criteria. SRCC quantifies the extent to which the ranks of two variables are related, which ranges from -1 to 1. Given  $N$  action images, SRCC is computed as:

$$SRCC = 1 - \frac{6 \sum_{n=1}^N (v_n - p_n)^2}{N(N^2 - 1)}, \quad (3)$$

where  $v_n$  and  $p_n$  denote the rank of the ground truth  $y_n$  and the rank of predicted score  $\hat{y}_n$  respectively. The higher the SRCC, the higher the monotonic correlation between ground truth and predicted score. Similarly, PLCC measures the linear correlation between predicted scores and ground truth scores, which can be formulated as:

$$PLCC = \frac{\sum_{n=1}^N (y_n - \bar{y})(\hat{y}_n - \bar{\hat{y}})}{\sqrt{\sum_{n=1}^N (y_n - \bar{y})^2} \sqrt{\sum_{n=1}^N (\hat{y}_n - \bar{\hat{y}})^2}}, \quad (4)$$

where  $\bar{y}$  and  $\bar{\hat{y}}$  are the mean of ground truth and predicted score respectively.



We also adopt the Kendall Rank Correlation Coefficient (KRCC) as an evaluation metric, which measures the ordinal association between two variables. For a pair of ranks  $(v_i, p_i)$  and  $(v_j, p_j)$ , the pair is concordant if:

$$(v_i - v_j)(p_i - p_j) > 0, \quad (5)$$

and discordant if  $< 0$ . Given  $N$  AIGVs, KRCC is computed as:

$$KRCC = \frac{C - D}{\frac{1}{2}N(N - 1)}, \quad (6)$$

where  $C$  and  $D$  denote the number of concordant and discordant pairs, respectively.

## 7.2. Detailed Information of Evaluation Methods

**LLaVA-1.5** [27] is an advanced Large Multimodal Model (LMM) framework designed for visual instruction tuning, aimed at improving multimodal understanding capabilities for general-purpose assistants. The model builds upon the LLaVA architecture and uses a simple fully-connected vision-language connector, making it more data-efficient.

**LLaVA-NeXT** [25] improves on LLaVA-1.5 [27] by increasing input image resolution and enhances visual detail, reasoning, and OCR capabilities. It also improves world knowledge and logical reasoning while maintaining LLaVA’s minimalist design and data efficiency, using under 1M visual instruction tuning samples.

**mPLUG-Owl3** [53] is a versatile multi-modal large language model designed to handle long image sequences, interleaved image-text, and lengthy video inputs. It introduces Hyper Attention blocks that efficiently integrate vision and language into a shared semantic space, allowing for the processing of extended multi-image scenarios.

**MiniCPM-V2.6** [52] is designed for deployment on end-side devices, addressing the challenges of running large models with significant computational costs. Key features include strong OCR capability, supporting high-resolution image perception, trustworthy behavior with low hallucination rates, and multilingual support for over 30 languages.

**Qwen2-VL** [41] is an advanced large vision-language model designed to process images, videos, and text with dynamic resolution handling and multimodal rotary position embedding (M-RoPE). The model features strong capabilities in OCR, video comprehension, multilingual support, and robust agent functionalities for device operations.

**Qwen2.5-VL** [2] is the latest flagship model in the Qwen vision-language series, featuring significant improvements in visual recognition, object localization, document parsing, and long-video comprehension. Building on the Qwen2-VL architecture, it introduces key enhancements such as dynamic resolution processing for images and videos, absolute time encoding for temporal dynamics, and window attention to optimize inference efficiency.

**Llama3.2-Vision** [31] excels in image reasoning tasks, such as document-level understanding, chart and graph captioning, and visual grounding. These models can reason with images, such as answering questions based on graphs or maps, and generate captions that describe visual scenes.

**DeepseekVL** [30] leverages a hybrid vision encoder for efficient high-resolution image processing and a carefully balanced training strategy that integrates language model capabilities with vision tasks. By emphasizing diverse, real-world data and a use case taxonomy, DeepSeek-VL delivers superior performance in tasks like OCR, document parsing, and visual-grounding.

**DeepseekVL2** [47] is an advanced series of mix-of-experts (MoE) vision language models. It introduces a dynamic tiling vision encoding strategy, allowing efficient processing of high-resolution images with varying aspect ratios, enhancing tasks like visual grounding and document analysis. It also leverages the Multi-head Latent Attention (MLA) mechanism for the language component, which reduces computational costs and improves inference efficiency.

**CogAgent** [17] is designed to facilitate understanding and navigation of graphical user interfaces (GUIs). It utilizes both low and high-resolution image encoders to recognize small text and page elements. CogAgent excels in GUI tasks like navigation and decision-making. CogAgent’s innovative design includes a cross-attention branch to balance high-resolution inputs and computational efficiency.

**InternVL2.5** [7] demonstrates strong performance in various benchmarks, including multi-discipline reasoning, document and video understanding, and multimodal hallucination detection. The model features enhanced vision encoders, larger dataset sizes, and improved test-time scaling.

**InternLM-XComposer** [54] excels at generating long-form content that integrates contextually relevant images, enhancing the engagement and immersion of the reading experience. It autonomously identifies optimal locations in the text for image placement and selects appropriate images from a large-scale database, ensuring contextual alignment.

**CLIPScore** [16] is an image captioning metric, which is widely used to evaluate T2I/T2V models. It passes both the image and the candidate caption through their respective feature extractors, then computing the cosine similarity between the text and image embeddings.

**BLIPScore** [26] provides more advanced multi-modal feature extraction capabilities. Using the same methodology as CLIPScore [16], it computes the cosine similarity between the text and visual embeddings, but benefits from enhanced pre-training strategy, which is designed to better capture fine-grained relationships between text and visual content.

**ImageReward** [51] builds upon the BLIP model [26] by introducing an additional MLP layer on top of BLIP’s output. Instead of directly computing a similarity score, the MLP

generates a scalar value representing the preference for one image over another in comparative settings.

**PickScore** [21] is a scoring function designed to predict human preferences in text-to-image generation. It was trained by fine-tuning CLIP-H on human preference data, aiming to maximize the probability of a preferred image being selected. PickScore exhibits strong correlation with human rankings, outperforming traditional metrics like FID and aesthetics predictors, and is recommended as a more reliable evaluation metric for text-to-image models.

**HPS** [45] is designed to improve text-to-image generation models by better aligning their outputs with human preferences. HPS is based on a fine-tuned CLIP model that accurately predicts human preferences over generated images.

**VQAScore** [23] is designed to assess the alignment between generated images and text prompts, particularly for compositional text-to-visual generation tasks. It can be used in a black-box manner, requiring no fine-tuning or additional prompt decomposition.

**FGA-BLIP2** [15] is a method for evaluating image-text alignment in T2I models, specifically designed to provide fine-grained analysis. It involves fine-tuning a vision-language model to produce alignment scores and element-level annotations for image-text pairs. This approach uses a variance-weighted optimization strategy to account for the diversity of images generated from specific prompts.

**CNNIQA** [19] is a convolutional neural network (CNN) designed for no-reference image quality assessment (NR-IQA), which predicts the visual quality of distorted images without using reference images. Unlike traditional methods that rely on handcrafted features, CNNIQA directly learns discriminative features from raw image patches, allowing for a more efficient and effective image quality estimation.

**DBCNN** [55] is a deep bilinear convolutional neural network designed for blind image quality assessment, which handles both synthetic and authentic distortions. The model uses two specialized convolutional neural networks. The features from both CNNs are pooled bilinearly into a unified representation for quality prediction.

**HyperIQA** [35] aims at handling authentically distorted images. It addresses two main challenges: distortion diversity and content variation. The model is based on a self-adaptive hyper network that adjusts quality prediction parameters according to the image content, making the predictions more consistent with human perception.

**TReS** [13] handles both synthetic and authentic distortions. It combines CNNs for capturing local image features with the self-attention mechanism to learn non-local features, addressing both local and global image quality aspects. The model also incorporates a relative ranking loss to enhance the correlation between subjective and objective scores by learning the relative quality ranking among images.

**MUSIQ** [20] leverages a patch-based multi-scale Trans-

former architecture to handle images of varying resolutions and aspect ratios without resizing or cropping. Unlike CNN-based models, which require fixed-size input, MUSIQ can process full-size images, extracting features at multiple scales to capture both fine-grained and global image quality details. The model introduces a unique hash-based 2D spatial embedding and scale embedding to effectively manage positional information across multi-scale inputs.

**StairIQA** [37] employs a staircase structure that hierarchically integrates features from intermediate layers of a CNN, allowing it to leverage both low-level and high-level visual information for more effective quality assessment. Additionally, it introduces an Iterative Mixed Database Training (IMDT) strategy, which trains the model across multiple diverse databases to improve generalization and handle variations in image content and distortions.

**Q-Align** [44] is a human-emulating syllabus designed to train large multimodal models for visual scoring tasks. It mimics the process of training human annotators by converting MOS into five text-defined rating levels. We used the officially pre-trained model and finetuned it on our EvalMi-50K.

**LIQE** [56] integrates auxiliary tasks such as scene classification and distortion type identification to improve the quality prediction of in-the-wild images. It uses a textual template to describe the image’s scene, distortion, and quality, using CLIP to compute the joint probability of these tasks.

### 7.3. Question design for LLM-based models

For LLM-based detection methods, we not only need to input the image to be evaluated, but also the corresponding prompt to guide the model to output the result we want. Three different questions need to be input for each image to be evaluated. When designing questions from the two dimensions of Perception and T2I correspondence, all images have a unified template, but to obtain the question-answer pair for an image, different questions need to be designed according to the challenge corresponding to the prompt used to generate the image. We have a total of 20 tasks, so there are 20 question models for this dimension. The specific question template is as follows:

- **Perception:** Suppose you are now a volunteer for subjective quality evaluation of images and you are now required to rate the quality of the given images on a scale of 0-100. Results are accurate to the nearest digit. Answer only one score.
- **T2I Correspondence:** Please rate the consistency between the image and the text description “<prompt>”. The rating scale is from 0 to 100, with higher scores for descriptions that include important content from the image and lower scores for descriptions that lack important content. Results are accurate to the nearest digit. Answer only a score.

• **Question-Answer Pairs:**

- (1) **Single class:** Does the image contain `<class_name >`? Answer yes or no.
- (2) **Two class:** Does the image contain both `<class1_name >` and `<class2_name >`? Answer yes or no.
- (3) **Counting:** Does the image contain `<class.count >``<class_name >`? Answer yes or no.
- (4) **Colors:** Does the image contain `<class_name >` in the color of `<class_color >`? Answer yes or no.
- (5) **Position:** Does the image contain both `<class1_name >` and `<class2_name >`, and are they positioned as described in “`<prompt >`”? Answer yes or no.
- (6) **Shapes:** Does the image contain a `<class.shape >``<class_name >`? Answer yes or no.
- (7) **Texture:** Does the image contain a `<class.texture >``<class_name >`? Answer yes or no.
- (8) **Scene:** Does the image depict a `<scene_name >` scene? Answer yes or no.
- (9) **Style:** Is the style of the image `<style_name >`? Answer yes or no.
- (10) **OCR (Optical Character Recognition):** Does the image contain the text “`<OCR >`” with all letters correct? Answer yes or no.
- (11) **HOI (Human-Object Interaction):** Does the image contain both a person and `<object_name >`, and is the person’s action `<verb_ing >`? Answer yes or no.
- (12) **Human:** Do the appearance, hairstyle, accessories, and profession of the person in the image match the description in “`<prompt >`”? Answer yes or no.
- (13) **Emotion:** If there is a person in the image, is their emotion `<emotion_class >`? If there is no person, does the overall mood of the image convey `<emotion_class >`? Answer yes or no.
- (14) **Linguistic Structure:** Does the scene depicted in the image exclude `<class_name >`? Answer yes or no.
- (15) **View:** Is the perspective shown in the image `<view_class >`? Answer yes or no.
- (16) **World Knowledge:** Does the image contain a famous landmark or celebrity `<knowledge_class >`? Answer yes or no.
- (17) **Face:** Does the face in the image have `<first.body_part >``<first.shape_or_color >` and `<second.body_part >``<second.shape_or_color >`? Answer yes or no.
- (18) **Imagination:** Does the image content show imaginative elements, and does it match the description in “`<prompt >`”? Answer yes or no.
- (19) **Time & Light:** Does the image depict the time `<time_class >` with sunlight appearing as `<light_class >`? Answer yes or no.
- (20) **Complex:** The questions for a complex challenge are a combination of the questions for the 19 individual challenges described above. For example, for a complex challenge consisting of a combination of task 1, task 2,

etc., the question template is: Are the text descriptions of the pictures: `<task1_question >`, `<task1_question >`... all correct? Answer yes or no.

The content in “`<>`” in the above question template needs to be determined based on the specific prompt content.

## 8. More Results Comparisons

As shown in Table 6, we further launch comparisons of the alignment between different metric results and human annotations in evaluating T2I model performance. We compare the performance of GenEval [12], Grounding-DINO [29], and our model across five tasks. Since GenEval [12] evaluates models using only these specific dimensions, we focus on tasks that align with GenEval’s capabilities to ensure a fair comparison. GenEval [12] evaluates object detection using Mask2Former [8], which is part of the MMDetection [6] toolbox from OpenMMLab, providing robust detection of objects and their relative positioning. For the counting task, Mask2Former [8] is paired with a higher confidence threshold (0.9) to improve human agreement. Additionally, a heuristic method is used to evaluate the relative positioning of objects based on their bounding box coordinates, classifying objects as “left”, “right”, “above”, or “below” one another if they meet a minimum distance threshold. For color classification, GenEval [12] utilizes the CLIP ViT-L/14 [33] model for zero-shot color classification, where each object’s bounding box is cropped to improve accuracy by removing the background.

To further explore the performance of detection models on these tasks, we replace Mask2Former [8] with Grounding-DINO [29] and use the InternVL2.5-38B [7] model for color classification. While this improves counting and position tasks due to Grounding-DINO’s enhanced detection, GenEval still outperforms on color, single-class, and two-class tasks. This is likely due to differences in detection model threshold settings and highlights the limitations of using detection models as a backbone for tasks such as counting and position, which may require more specialized methods. In contrast, our model, which combines LMM for comprehensive evaluation, outperforms both GenEval [12] and Grounding-DINO [29] in all tasks. Unlike GenEval [12], which relies on a combination of multiple models to handle different tasks, our approach is an *all-in-one* solution that integrates various capabilities into a single framework. This unified design allows for more consistent and efficient performance across tasks, as it avoids the potential inconsistencies and complexities that arise from combining multiple specialized models. Our model demonstrates superior task-specific accuracy, achieving higher human agreement and better overall performance across all tasks, showcasing the advantage of our integrated approach over traditional detection-based methods and multi-model systems.

Table 3. Performance comparisons of T2I Models on human-annotated perception MOS.

Models	Single	Two Class	Counting	Colors	Position	Shapes	Texture	Scene	Style	OCR	HOI	Human	Emotion	Linguistic	View	Knowledge	Face	Imagination	Time&Light	Complex	Overall	Rank
Playground [24]	63.56	61.78	62.20	64.19	58.84	62.86	63.40	63.34	61.98	55.54	61.09	63.80	61.74	60.29	58.96	61.77	59.66	61.97	61.82	61.76	61.64	1
Kolours [40]	63.47	61.51	61.58	63.96	59.59	61.92	61.93	61.40	62.53	53.01	59.75	62.50	62.18	59.73	59.34	61.64	59.27	60.67	60.96	61.47	61.14	2
Infinity [14]	65.31	60.68	61.67	65.02	58.02	60.60	62.90	63.76	61.10	64.78	59.73	61.82	60.54	59.26	57.81	58.34	56.53	61.67	61.65	60.86	1	3
Flux.schnell [22]	65.17	62.99	61.82	63.05	59.66	62.46	63.28	65.45	57.71	65.83	63.31	62.78	60.27	59.50	58.92	58.96	48.76	59.31	52.26	63.05	60.63	4
SD3.5_large [10]	64.37	62.48	61.71	63.93	58.39	60.70	61.62	57.26	59.57	65.27	56.56	60.71	57.18	57.64	57.29	57.47	49.87	60.48	52.95	62.40	59.50	5
DALLE3 [3]	63.01	62.72	60.32	62.09	57.93	60.79	63.32	55.06	63.46	67.73	58.76	59.88	59.22	56.37	58.05	57.60	45.88	57.82	54.96	61.81	59.34	6
OmniGen [49]	63.47	60.03	59.28	61.53	55.72	59.13	60.02	60.25	57.98	57.82	58.07	63.87	58.90	57.22	56.79	58.87	60.89	57.65	55.48	59.11	59.12	7
Kandinsky-3 [11]	59.78	55.50	58.03	60.32	54.19	60.74	60.47	60.47	61.56	52.61	58.25	58.83	57.24	56.92	56.24	57.56	62.60	56.73	58.74	57.53	58.21	8
PixArt-sigma [5]	60.89	56.52	57.87	58.88	53.52	59.72	60.37	60.26	59.19	48.37	55.30	58.85	57.35	57.06	56.67	56.46	54.27	59.71	59.91	56.57	57.43	9
EMU3 [42]	57.08	53.58	53.53	54.56	50.78	54.74	55.73	57.55	57.23	43.02	53.72	57.73	54.56	54.37	52.81	54.83	56.08	54.19	55.81	52.72	54.29	10
SDXL_base-1 [32]	59.34	56.33	57.49	59.84	52.48	56.53	57.72	51.87	54.79	49.90	50.76	53.56	49.53	50.07	53.05	54.38	41.55	52.46	50.96	54.81	53.50	11
Show-o [50]	60.81	57.33	59.30	60.59	53.37	58.74	60.58	52.74	53.91	41.50	47.96	46.43	45.46	50.20	50.35	51.64	37.21	52.20	45.91	54.53	52.31	12
Seed-xi [11]	55.06	46.23	52.50	54.60	45.34	55.15	55.74	52.62	53.99	49.19	46.27	47.45	50.63	49.20	52.09	54.50	42.58	53.15	52.10	49.97	50.73	13
NOVA [9]	56.81	54.23	52.95	57.69	50.65	54.41	56.36	49.77	57.76	31.76	45.43	43.87	47.44	48.77	46.98	49.36	48.58	53.43	47.53	50.85	50.69	14
LaVi-Bridge [57]	56.13	52.18	52.74	54.03	45.96	54.44	54.04	52.53	56.12	39.37	51.42	46.85	50.60	50.38	48.25	48.11	45.04	51.62	50.38	49.95	50.56	15
Hart [38]	52.12	48.76	49.54	53.19	46.97	50.65	51.14	47.09	52.50	39.89	42.12	50.06	50.95	48.31	49.08	50.27	53.21	53.99	53.90	48.06	49.80	16
LLMGA [48]	53.30	52.14	52.41	54.92	47.61	54.17	54.94	50.95	53.03	50.28	43.30	42.90	46.14	49.46	49.40	49.74	39.68	47.69	49.26	44.50	48.67	17
SD_v2-1 [34]	55.85	49.76	53.15	56.41	44.87	52.90	51.38	48.46	43.40	42.72	44.30	47.00	42.28	48.90	50.17	50.99	35.35	39.82	47.13	48.87	47.68	18
ELLA [18]	48.78	46.21	46.21	50.75	43.48	49.55	52.33	41.93	40.25	38.11	41.99	43.38	42.24	42.70	43.44	40.32	31.11	42.76	45.97	49.63	44.61	19
Janus [43]	42.57	40.18	37.82	41.99	36.58	41.00	41.06	38.91	40.53	26.47	33.63	30.69	33.34	34.77	36.89	37.97	29.81	34.26	40.89	37.24	36.98	20
i-Code-V3 [39]	42.58	36.07	37.27	37.96	30.44	40.41	39.92	35.61	38.71	33.44	32.81	30.63	29.93	36.84	32.27	32.23	34.45	29.49	35.50	33.48	34.70	21
Vila-u [46]	38.74	32.54	33.44	38.15	29.88	38.00	35.26	33.24	40.05	27.61	28.72	27.20	32.37	33.12	32.64	34.38	37.29	34.44	39.99	31.38	33.86	22
LlamaGen [36]	33.86	30.90	33.12	33.96	27.89	34.17	35.17	32.53	33.29	27.81	29.46	29.05	26.52	33.19	29.34	32.96	21.72	28.77	27.59	27.04	29.96	23
LWM [28]	35.11	29.55	32.08	32.68	25.82	36.12	33.10	30.87	30.95	34.21	29.23	24.33	24.15	30.64	26.00	26.17	29.18	22.50	29.44	26.89	28.88	24

Table 4. Performance comparisons of T2I Models on human-annotated correspondence MOS.

Models	Single	Two Class	Counting	Colors	Position	Shapes	Texture	Scene	Style	OCR	HOI	Human	Emotion	Linguistic	View	Knowledge	Face	Imagination	Time&Light	Complex	Overall	Rank
SD3.5_large [10]	64.96	62.10	60.32	63.79	45.58	55.56	60.48	64.10	60.59	65.66	60.20	62.59	58.47	38.73	56.59	61.72	55.66	59.39	58.17	57.75	58.35	1
Flux.schnell [22]	64.88	63.19	58.71	58.32	49.32	53.16	57.91	66.11	59.31	65.29	62.68	62.64	59.72	35.54	58.40	62.69	56.94	61.40	56.68	56.03	58.10	2
DALLE3 [3]	64.57	63.50	55.28	63.60	48.29	54.85	58.75	62.87	57.25	63.84	61.07	62.64	61.16	40.73	60.85	62.67	53.79	62.01	59.34	53.22	57.97	3
Infinity [14]	65.42	57.82	56.31	63.99	45.19	48.36	54.79	65.92	61.22	60.64	59.73	62.46	59.23	35.71	57.32	62.65	58.95	59.66	61.75	55.56	57.43	4
Playground [24]	63.59	59.00	55.04	62.99	39.47	54.98	58.50	64.98	60.74	38.42	60.42	62.27	60.67	42.52	57.39	63.72	59.77	58.87	62.38	46.53	56.06	5
OmniGen [49]	64.09	58.38	50.47	60.11	46.09	50.75	51.29	65.97	55.68	55.58	58.72	61.79	58.15	39.50	56.32	62.32	61.13	57.04	60.60	50.59	55.81	6
PixArt-sigma [5]	62.22	52.68	52.77	60.94	40.56	50.34	59.35	64.25	62.16	31.97	57.76	62.90	58.27	38.33	56.96	61.65	58.70	59.68	62.06	46.65	54.72	7
Show-o [50]	63.12	57.86	59.20	62.27	44.73	52.25	55.37	63.32	57.28	30.69	56.10	58.04	54.03	39.29	55.75	59.03	50.80	56.71	55.07	50.66	54.21	8
Kolours [40]	63.71	56.02	53.28	59.78	39.79	51.03	50.65	62.75	40.29	39.84	54.83	60.36	58.42	35.45	55.44	62.09	55.32	56.97	61.45	46.01	53.53	9
NOVA [9]	59.65	55.74	53.29	60.80	40.38	53.12	55.61	61.70	58.29	27.00	56.62	57.01	55.00	38.47	54.32	58.26	55.66	56.33	54.55	45.99	52.73	10
SDXL_base-1 [32]	61.38	52.38	48.85	60.34	39.16	50.52	54.82	62.24	59.17	44.22	57.69	58.78	53.92	41.05	54.75	60.28	49.12	53.54	56.90	42.91	52.23	11
EMU3 [42]	58.40	48.65	44.81	57.77	38.21	47.31	50.93	63.35	56.42	32.13	55.50	59.78	55.39	38.94	54.04	59.47	55.87	54.61	59.02	40.79	50.97	12
Seed-xi [11]	58.52	46.69	44.94	59.55	39.25	51.09	53.58	63.53	57.93	34.77	55.63	53.93	55.31	39.40	55.25	59.51	48.98	56.33	57.77	41.32	50.96	13
Hart [38]	55.22	45.66	45.89	57.49	38.51	47.00	51.51	59.73	55.46	27.44	51.14	57.91	55.19	38.94	53.85	56.30	57.59	54.85	59.83	41.71	50.30	14
LaVi-Bridge [57]	60.12	50.33	49.29	59.82	34.70	48.43	52.16	63.31	57.69	27.12	56.24	56.95	54.66	37.88	52.94	54.45	50.32	53.63	58.55	39.07	50.19	15
ELLA [18]	56.15	46.30	47.71	58.26	38.95	48.10	52.11	60.67	49.89	31.52	51.98	56.18	51.15	36.08	51.61	53.65	41.37	50.66	53.69	46.17	49.07	16
Kandinsky-3 [11]	58.71	42.14	46.52	53.22	34.37	50.46	45.89	58.91	50.42	30.55	52.94	55.30	54.10	36.87	52.78	58.61	56.54	54.77	57.56	35.43	48.37	17
SD_v2-1 [34]	59.86	46.84	46.62	58.48	34.31	49.30	50.78	60.46	52.24	34.64	53.37	54.08	47.89	44.59	54.01	56.84	42.22	42.88	50.81	38.50	47.96	18
Janus [43]	50.90	44.38	39.35	56.09	46.98	42.13	43.46	58.68	51.88	26.38	44.85	49.30	45.30	40.04	50.66	50.96	43.03	42.27	50.38	42.06	45.94	19
Vila-u [46]	47.69	37.88	39.35	51.55	33.31	43.94	41.47	51.54	50.91	26.34	43.77	48.24	46.44	40.08	48.16	48.69	49.04	43.81	53.18	34.68	43.47	20
LLMGA [48]	53.92	40.69	36.84	49.10	32.85	41.44	41.88	60.43	50.41	38.52	42.97	41.36	46.74	43.92	47.98	51.45	42.37	46.45	51.73	32.54	43.43	21
i-Code-V3 [39]	49.24	34.35	39.11	48.28	28.87	41.66	42.94	56.29	47.69	29.26	43.76	44.32	38.01	40.94	42.70	42.83	39.39	33.45	43.31	30.85	39.80	22
LlamaGen [36]	44.06	35.45	37.81	46.75	31.34	38.34	40.49	49.64	44.38	27.37	43.08	43.30	36.59	38.08	43.23	44.96	27.54	34.75	35.42	29.77	37.73	23
LWM [28]	43.55	32.29	34.56	43.80	28.10	40.23	36.69	46.71	42.15	35.79	38.64	35.38	32.62	36.44	37.06	35.08	36.62	28.22	35.75	29.35	35.46	24

Table 5. Performance comparisons of T2I Models on human-annotated task-specific accuracy.

Models	Single	Two Class	Counting	Colors	Position	Shapes	Texture	Scene	Style	OCR	HOI	Human	Emotion	Linguistic	View	Knowledge	Face	Imagination	Time&Light	Complex	Overall	Rank
SD3.5_large [10]	98.89	94.50	82.22	91.35	36.94	68.75	86.00	96.97	88.66	94.00	91.86	98.10	90.09	23.81	75.25	97.00	76.79	82.88	87.13	78.42	81.43	1
Flux.schnell [22]	95.56	94.50	74.44	76.92	48.65	60.00	77.00	98.48	89.69	94.00	95.35	95.24	90.99	11.90	84.16	100.00	83.04	96.40	83.17	71.92	80.29	2
DALLE3 [3]	100.00	94.50	66.67	93.27	47.75	70.00	83.00	95.45	73.20	90.00	89.53	97.14	95.50	29.76	92.08	97.00	70.54	94.90	90.10	64.73	80.24	3
Infinity [14]	100.00	78.90	67.78	90.38	37.84	47.50	67.00	100.00	92.78	80.00	89.53	97.14	89.19	13.10	83.17	99.00	83.93	88.29	93.07	71.23	78.10	4
Playground [24]	94.44	84.40	65.56	94.23	22.52	68.75	78.00	98.48	88.66	12.00	89.53	97.14	95.50	33.33	82.18	100.00	86.61	85.59	93.07	41.10	73.86	5
OmniGen [49]	96.67	83.49	50.00	79.81	39.64	56.25	55.00	100.00	75.26	68.00	87.21	97.14	84.68	23.81	78.22	97.00	91.96	78.38	93.07	56.51	73.29	6
Show-o [50]	96.67	82.57	78.89	89.42	36.94	58.75	70.00	95.45	80.00	82.56	93.33	79.28	25.00	80.20	93.00	63.39	83.78	80.20	57.88	71.71	7	
PixArt-sigma [5]	88.89	62.39	60.00	86.54	23.42	50.00	81.00	93.94	96.91	2.00	82.56	99.05	85.59	21.43	84.16	97.00	80.50	88.29	96.04	43.15	70.71	8
NOVA [9]	91.11	76.15	64.44	89.42	23.42	67.50	72.00	98.48	84.54	81.20	82.56	94.00	79.28	20.24	78.22	99.00	87.36	82.88	90.10	41.78	68.87	9
FLUX.1 [1]	90.00	71.56	60.00	84.22	23.42	53.75	65.00	90.91	74.23	14.00	88.66	89.53	88.25	13.10	84.68	97.00	87.38	92.08	93.07	65.05	10	
SDXL-base-1 [32]	92.22	62.22	43.33	87.50	20.72	56.25	69.00	93.94	89.69	30.00	83.72	92.38	72.97	28.57	72.31	93.00	46.43	63.96	82.18	33.56	63.67	11
Seedix [11]	87.78	42.20	34.44	87.50	16.22	61.25	65.00	100.00	87.63	4.00	87.21	78.10	83.78	25.00	79.23	97.00	49.41	82.88	87.13	27.05	61.43	12
EMU3 [42]	83.33	47.71	34.44	80.77	16.22	41.25	53.00	96.97	79.38	2.00	75.58	91.43	77.48	25.00	71.29	96.00	77.68	73.87	89.11	25.68	59.90	13
Hart [38]	72.22	38.53	38.89	75.00	19.82	38.75	56.00	87.88	79.38	0.00	61.63	84.76	84.68	25.00	78.22	86.00	83.04	71.17	94.06	31.51	59.29	14
LaVi-Bridge [57]	87.78	59.63	50.00	87.50	9.91	51.25	56.00	96.97	79.38	0.00	79.07	85.71	78.38	20.24	74.26	75.00	57.14	68.47	91.09	23.29	59.10	15
ELLA [18]	75.56	42.20	46.67	83.65	19.82	52.50	61.00	96.97	58.76	4.00	61.63	84.76	66.67	16.67	64.36	73.00	33.04	51.35	68.32	44.86	54.90	16
Kandinsky-3 [1]	81.11	25.69	41.11	61.54	9.01	52.50	38.00	81.82	48.45	0.00	60.47	74.29	70.27	16.67	69.31	91.00	73.21	70.27	79.21	12.67	50.14	17
SD_v2-1 [34]	88.89	46.79	41.11	82.69	8.11	53.75	52.00	89.39	62.89	4.00	69.77	74.29	39.64	42.86	75.25	87.00	25.00	19.82	51.49	21.58	48.86	18
Janus [43]	61.11	29.36	18.89	77.88	51.35	26.25	23.00	87.88	65.98	0.00	38.37	54.29	31.53	26.19	67.33	63.00	34.82	17.12	59.41	33.56	42.95	19
LLMGA [48]	70.00	28.44	13.33	51.92	4.50	30.00	25.00	87.88	59.79	26.00	34.88	37.14	50.45	39.29	54.46	68.00	25.89	40.54	62.38	10.27	37.67	20
Vila-u [46]	48.89	11.01	27.78	62.50	9.01	35.00	19.00	56.06	63.92	0.00	32.56	50.48	36.04	30.95	53.47	49.00	47.32	19.82	71.29	14.04	35.24	21
SD-Text-V3 [39]	60.00	8.26	22.22	54.81	0.00	8.75	34.00	78.79	47.42	2.00	26.74	36.19	11.71	34.52	32.67	29.00	14.64	3.60	23.76	5.48	25.00	22
LlamaGen [36]	36.67	8.26	18.89	51.92	8.11	16.25	27.00	46.97	32.99	0.00	36.05	32.38	11.71	25.00	42.57	39.00	0.00	5.41	12.87	6.85	21.19	23
Llama [28]	36.67	1.83	11.11	38.46	1.80	25.00	14.00	53.03	24.74	18.00	12.79	17.14	3.60	17.86	25.74	18.00	12.50	0.00	14.85	5.14	15.48	24



Perception 4-5 (**Excellent**): The image is nearly flawless, with high detail, accurate colors, and no visible artifacts, achieving professional-quality standards.



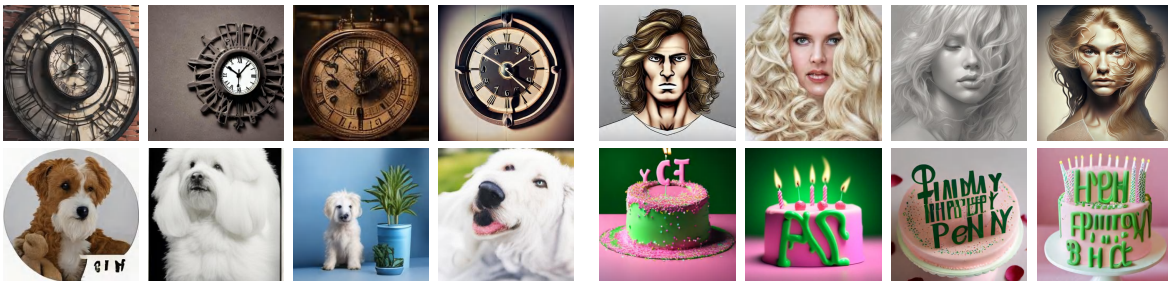
Perception 3-4 (**Good**): The image is visually appealing with minor flaws, offering clear details and natural colors, suitable for most applications.



Perception 2-3 (**Fair**): The image is somewhat acceptable but contains noticeable imperfections, such as mild artifacts or features that clearly indicate it is AI-generated.



Perception 1-2 (**Poor**): The image has significant flaws, such as heavy artifacts, poor detail, or unnatural colors, making it visually unappealing.



Perception 0-1 (**Bad**): The image is severely distorted, unrecognizable, or fails to convey any visual information.



Figure 5. Instructions and examples for manual evaluation of **perception**.



Correspondence 4-5 (**Excellent**): The image perfectly matches the text, capturing all details, relationships, and nuances.



Correspondence 3-4 (**Good**): The image closely aligns with the text, accurately representing most described elements with minor errors or omissions.



Perception 2-3 (**Fair**): The image partially matches the text but has significant inconsistencies, such as missing key objects or incorrect attributes.



Perception 1-2 (**Poor**): The image shows minimal alignment with the text, containing incorrect representations of the described elements.



Perception 0-1 (**Bad**): The image completely fails to match the text description.

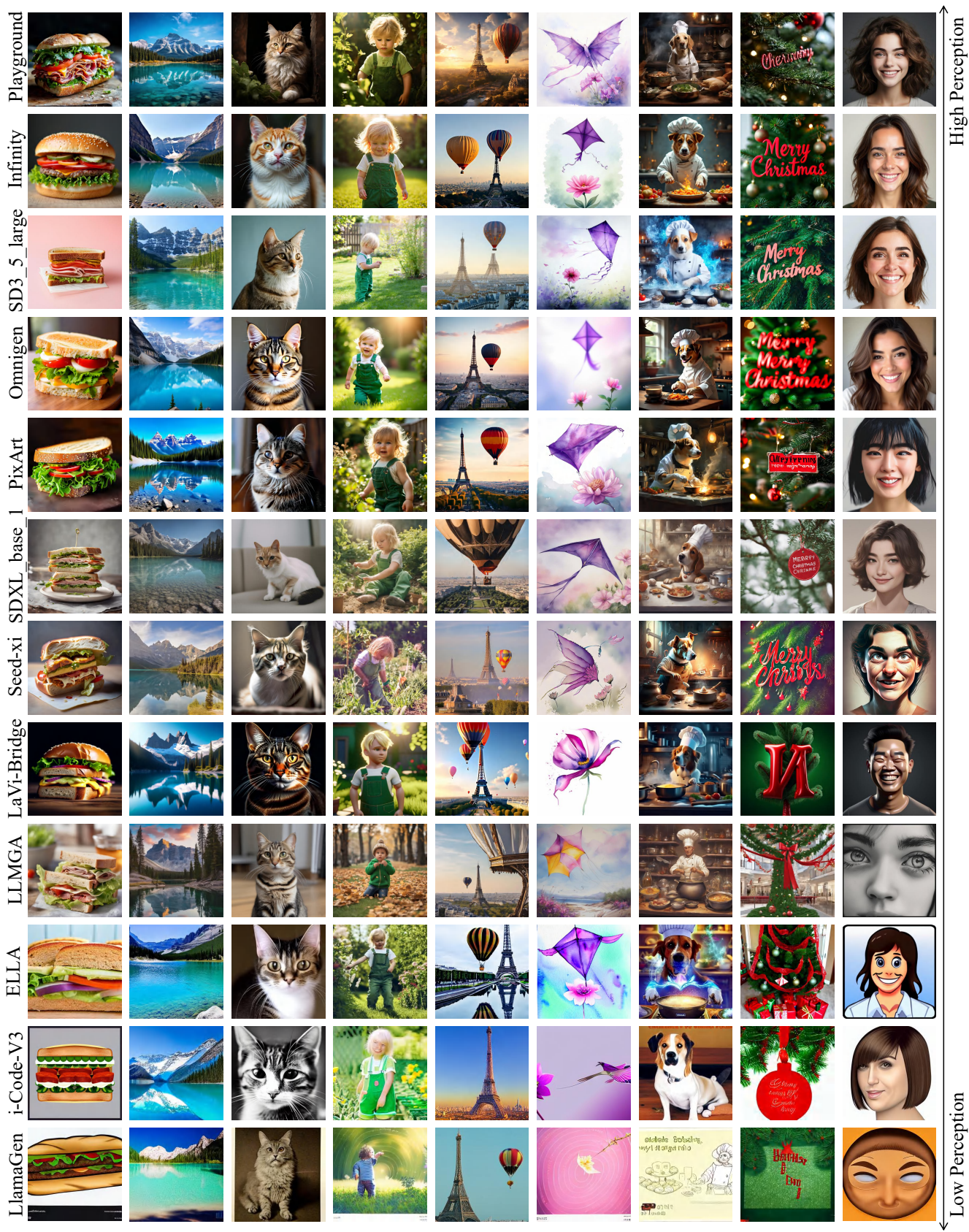


Yellow phrase "Best Wishes" on a blue box.

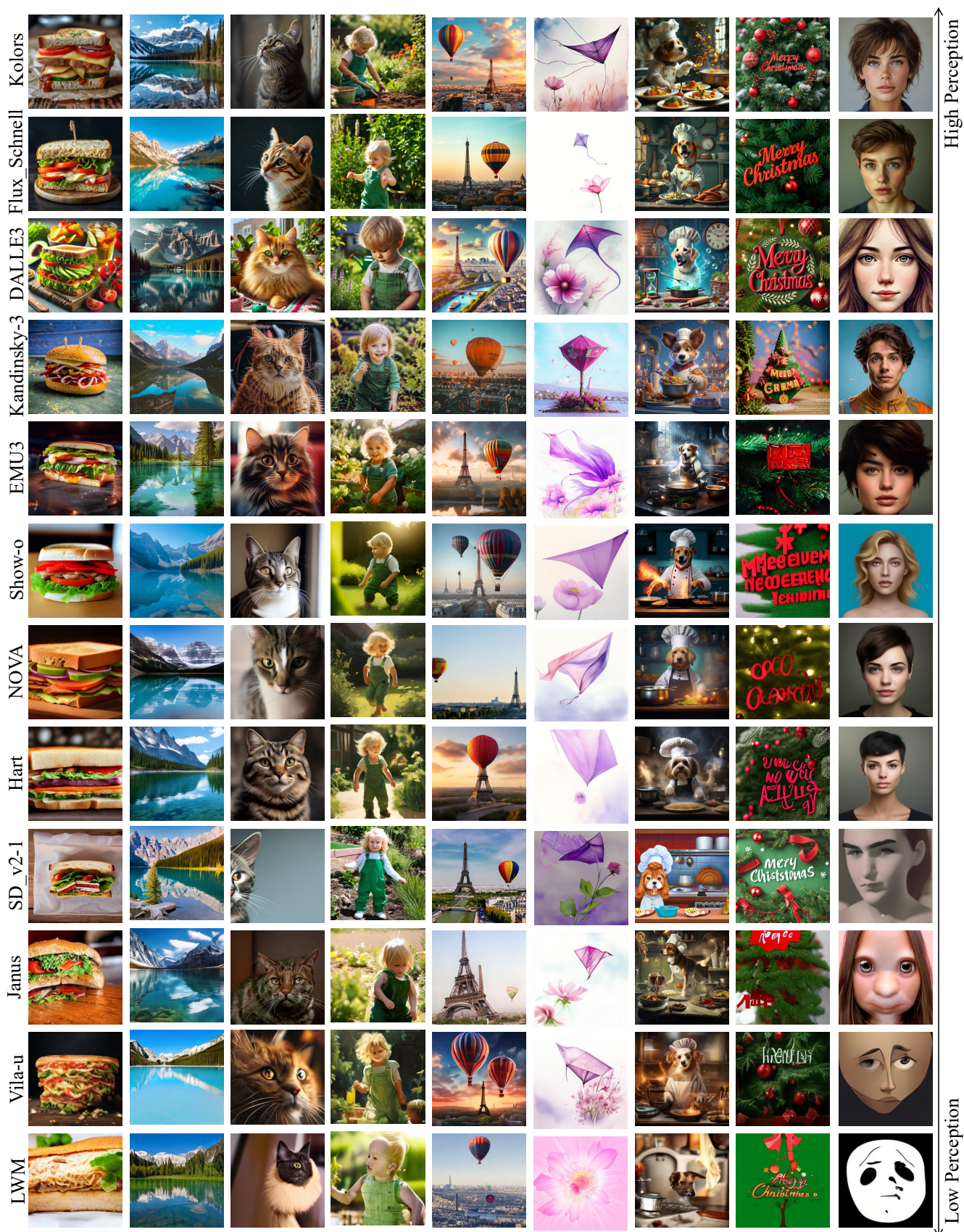
Two blue dogs and three black cats.

Figure 6. Instructions and examples for manual evaluation of **T2I correspondence**. Prompt (left): yellow phrase "Best Wishes" on a blue box. Prompt (right): two blue dogs and three black cats.











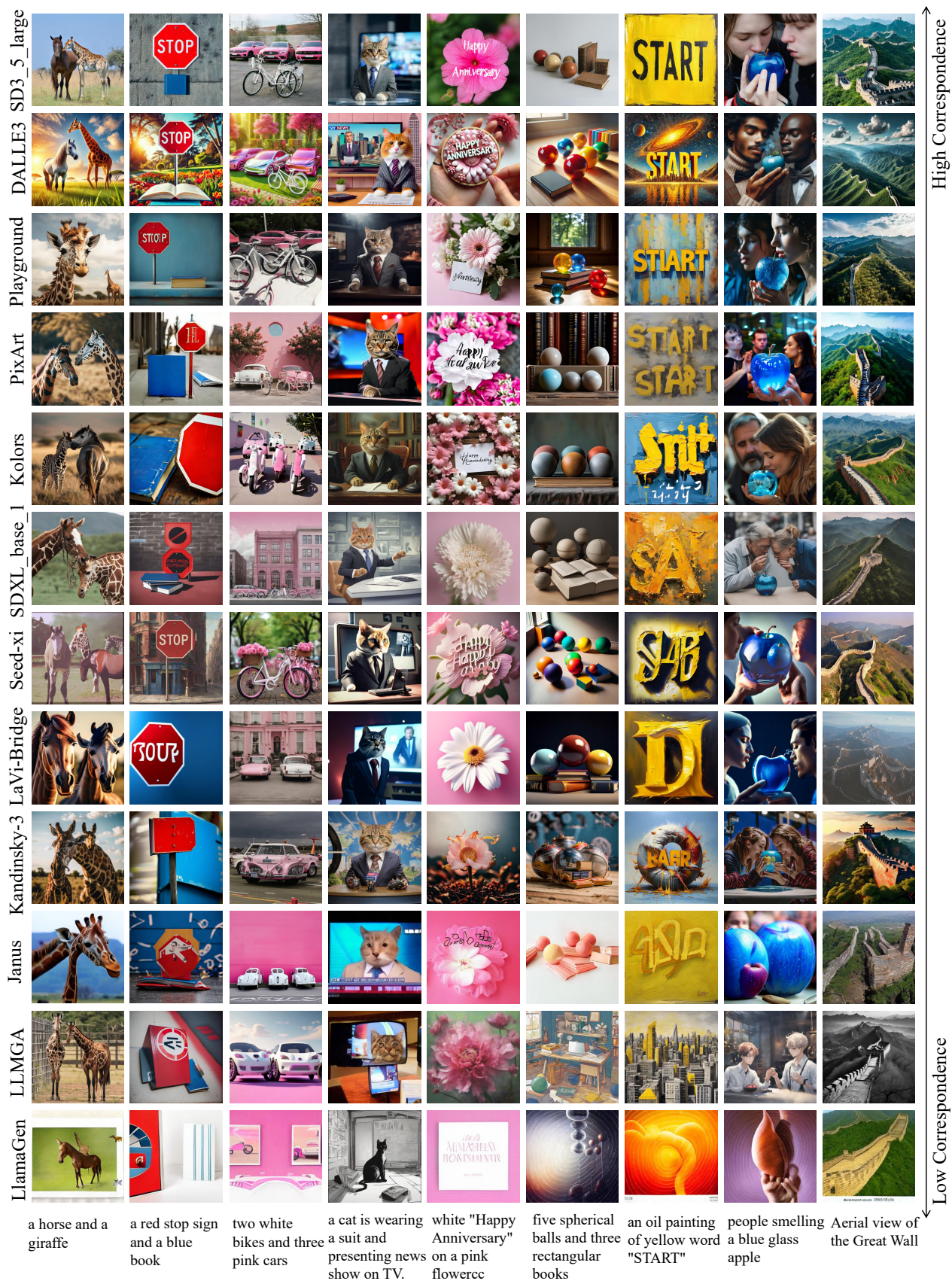


Figure 9. Visualization of generated videos in the EvalMi-50K: sort by average **T2I correspondence** of T2I models from high to low.



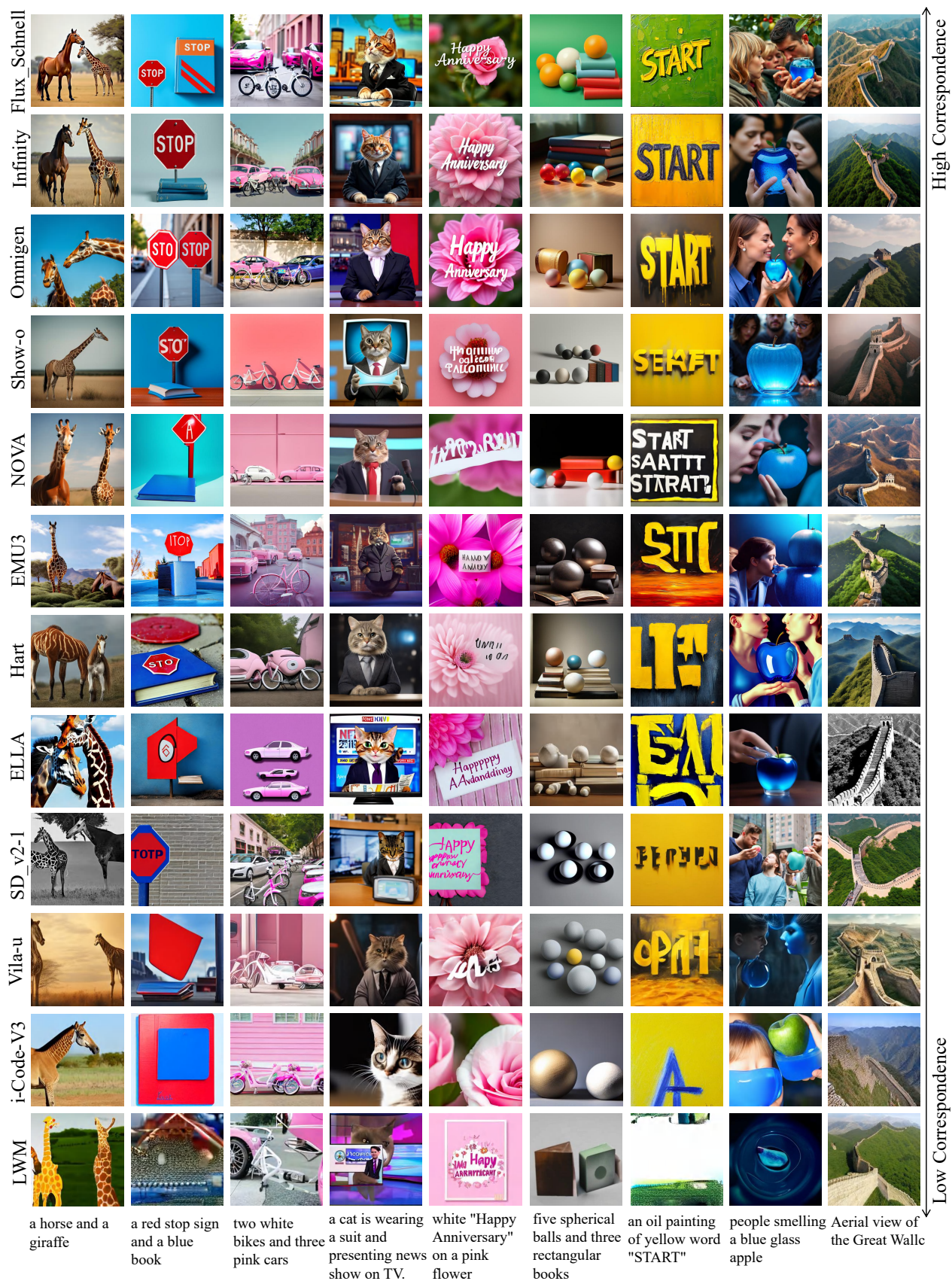
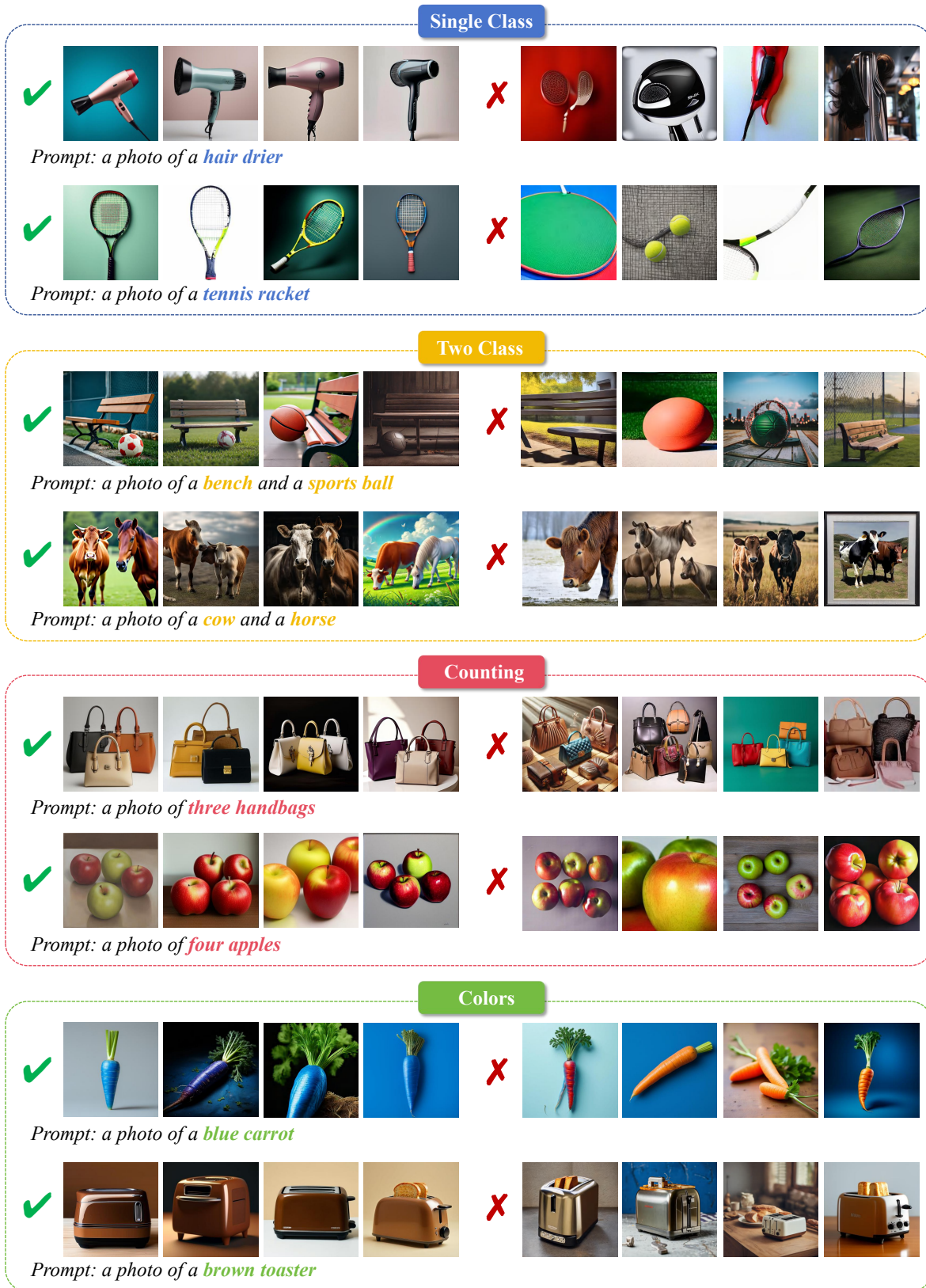
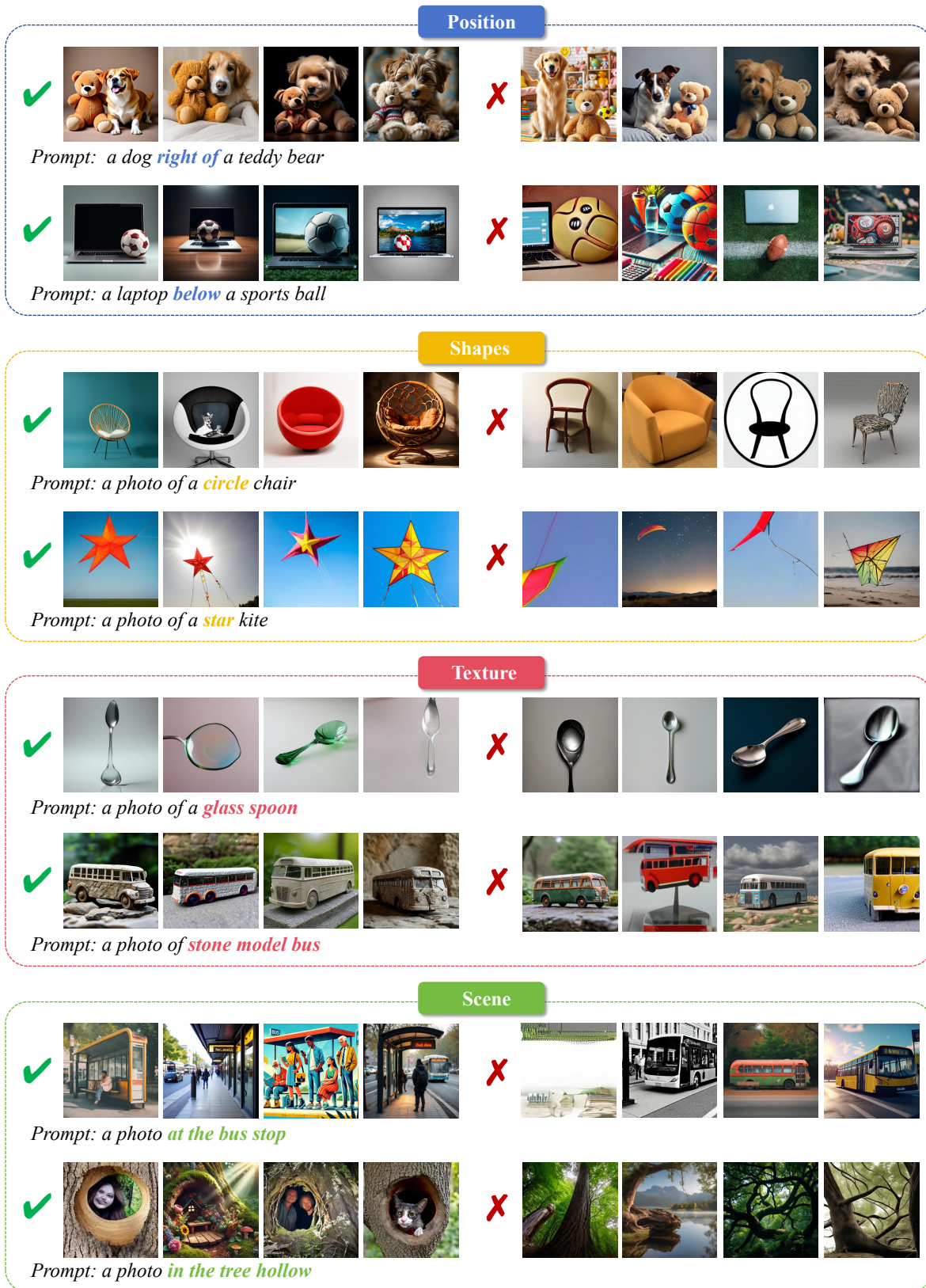


Figure 10. Visualization of generated videos in the EvalMi-50K: sort by average **T2I correspondence** of T2I models from high to low.









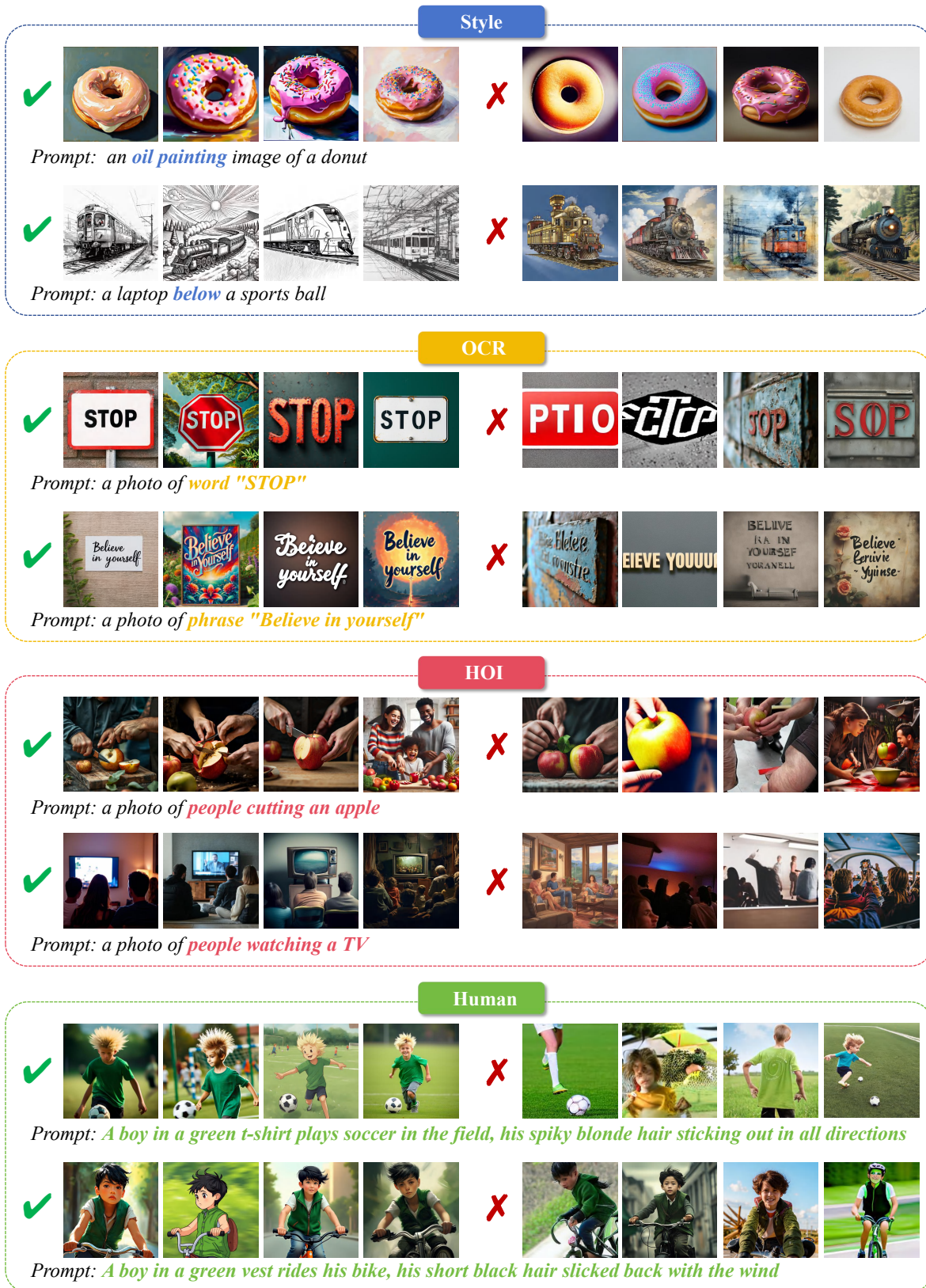


Figure 13. Examples for different task-specific challenges.

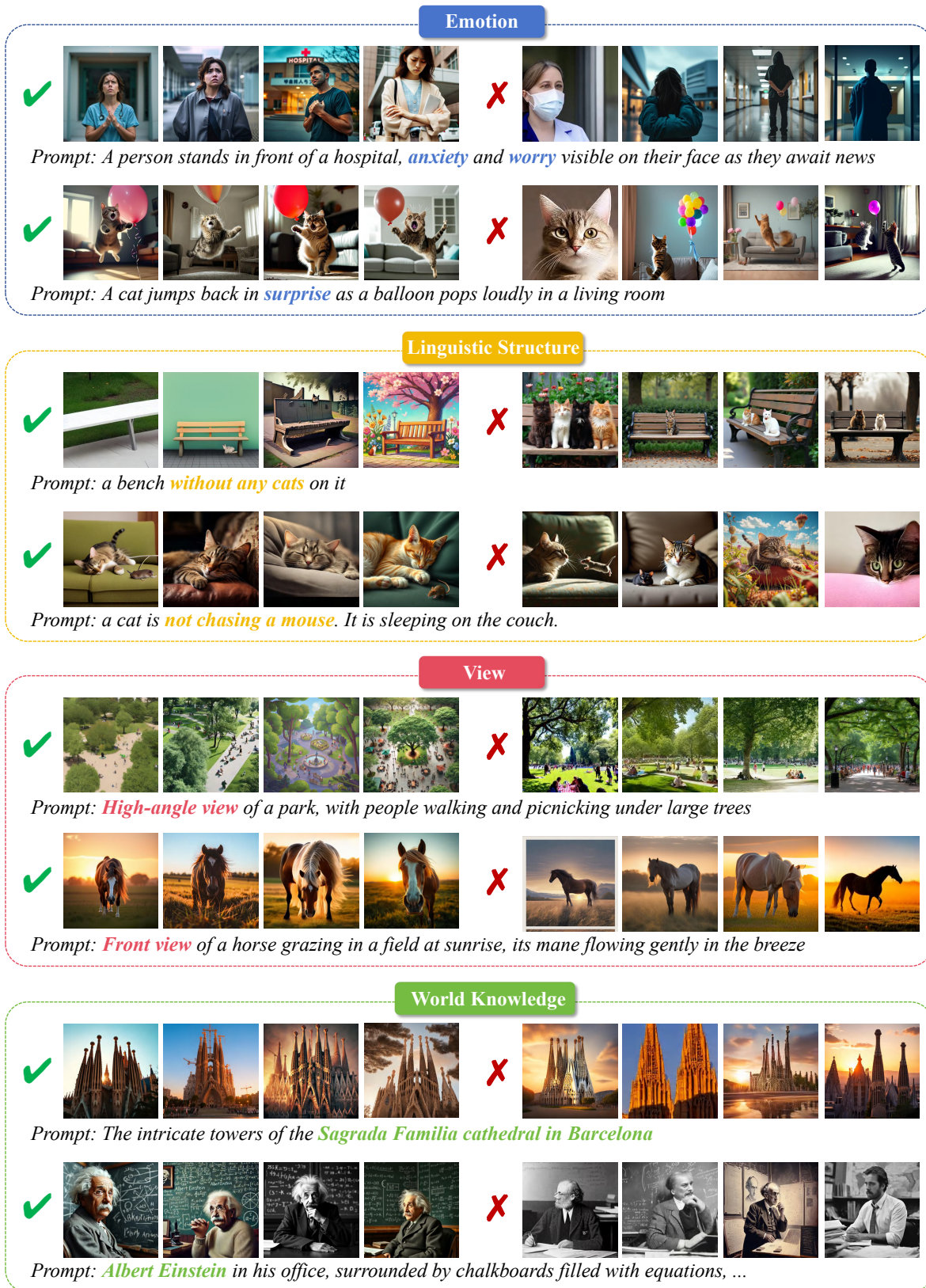


Figure 14. Examples for different task-specific challenges.



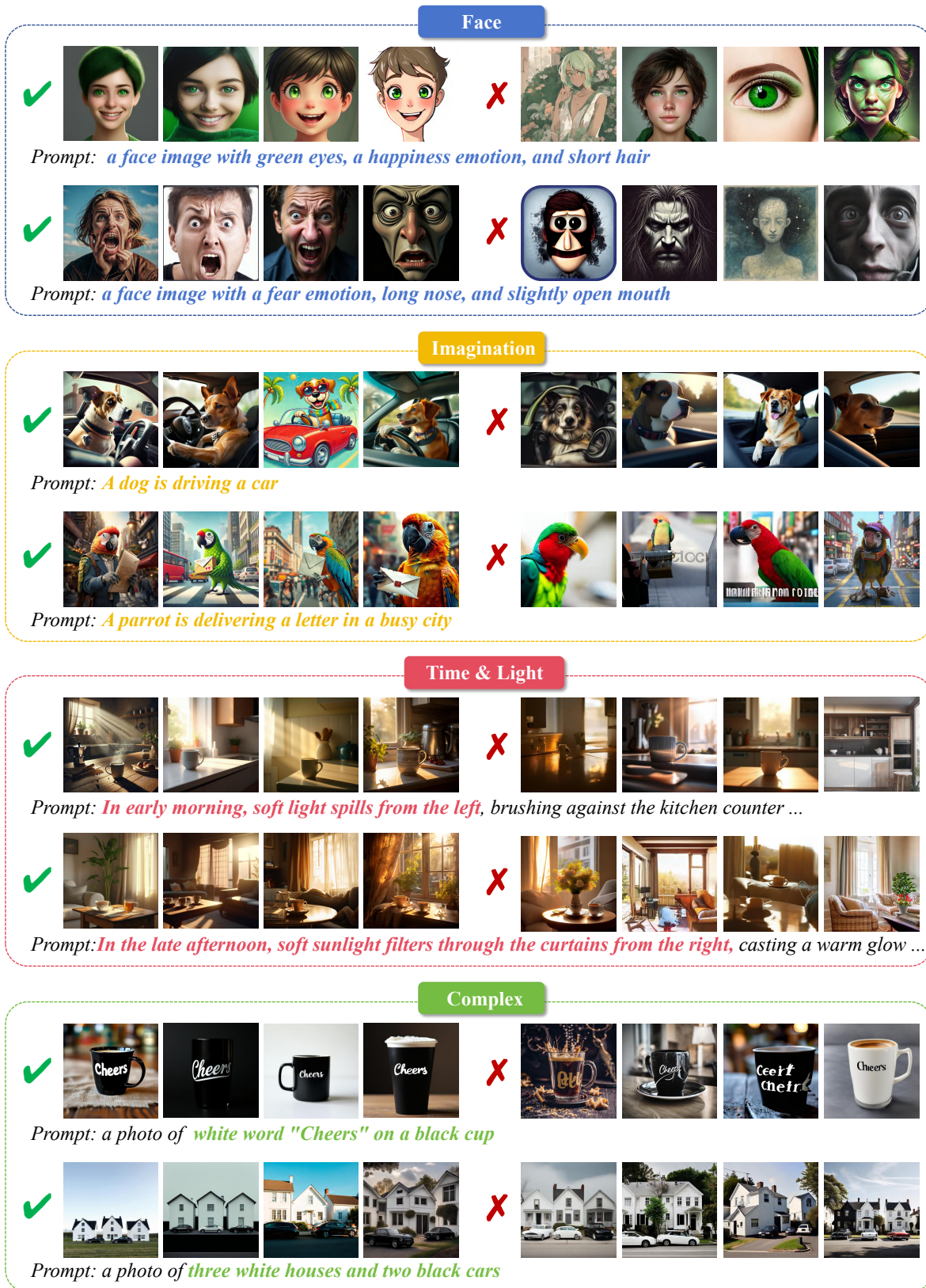


Figure 15. Examples for different task-specific challenges.



## References

- [1] Vladimir Arkhipkin, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Anton Bukashkin, Konstantin Kulikov, et al. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. *arXiv preprint arXiv:2410.21061*, 2024. 3, 4, 13
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 10
- [3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. 3, 4, 13
- [4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 4
- [5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\sigma$ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 74–91, 2024. 4, 13
- [6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 12
- [7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 10, 12
- [8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 12
- [9] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 4, 5, 13
- [10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 4, 8, 13
- [11] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 4, 13
- [12] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 52132–52152, 2023. 12
- [13] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3209–3218, 2022. 11
- [14] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bit-wise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 4, 5, 13
- [15] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *arXiv preprint arXiv:2412.18150*, 2024. 11
- [16] 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. 10
- [17] Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2024. 10
- [18] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 3, 4, 13
- [19] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740, 2014. 11
- [20] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5148–5157, 2021. 11
- [21] 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. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 36652–36663, 2023. 11
- [22] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 4, 13

- [23] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5290–5301, 2024. 11
- [24] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 3, 4, 8, 13
- [25] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 10
- [26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International conference on machine learning (ICML)*, pages 12888–12900, 2022. 10
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 10
- [28] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024. 3, 4, 13
- [29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–55, 2024. 12
- [30] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 10
- [31] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024, 2024. 10
- [32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 4, 13
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning (ICML)*, pages 8748–8763. PmLR, 2021. 3, 12
- [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3, 4, 13
- [35] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3667–3676, 2020. 11
- [36] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 4, 13
- [37] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 17(6):1178–1192, 2023. 9, 11
- [38] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 4, 5, 13
- [39] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 16083–16099, 2023. 3, 4, 13
- [40] Kolrs Team. Kolrs: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024. 4, 8, 13
- [41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 10
- [42] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 4, 13
- [43] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 4, 5, 13
- [44] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching llms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 9, 11
- [45] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2096–2105, 2023. 11

- [46] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. [4](#), [13](#)
- [47] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. [10](#)
- [48] Bin Xia, Shiyin Wang, Yingfan Tao, Yitong Wang, and Jiaya Jia. Llmga: Multimodal large language model based generation assistant. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 389–406, 2024. [3](#), [4](#), [13](#)
- [49] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. [4](#), [13](#)
- [50] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. [4](#), [5](#), [13](#)
- [51] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 15903–15935, 2023. [10](#)
- [52] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. [10](#)
- [53] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multimodal large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. [10](#)
- [54] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. [10](#)
- [55] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(1):36–47, 2018. [11](#)
- [56] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14071–14081, 2023. [9](#), [11](#)
- [57] Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K Wong. Bridging different language models and generative vision models for text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2024. [3](#), [4](#), [13](#)