MMGenBench: An Automated Pipeline for Evaluating LMMs from a Text-to-Image Generation Perspective

Anonymous ACL submission

Abstract

Large Multimodal Models (LMMs) demonstrate impressive capabilities. However, many existing benchmarks predominantly focus on image comprehension in specific domains and are labor-intensive to construct; their reference answers also tend to be brief, limiting their ability to assess whether LMMs can generate detailed image descriptions. To address these limitations, we propose MMGenBench-Pipeline, an automated evaluation pipeline that elicits textual descriptions from input images, synthesizes auxiliary images from these descriptions via text-to-image generative models, and compares the original and generated images at the representation level. We further construct MMGenBench-Domain, which targets the generated-image domain and demonstrates domain-specific evaluation of LMMs by solely utilizing image inputs, and MMGenBench-Test, which probes the fundamental image understanding and description capabilities of LMMs across 13 distinct image patterns. Experiments on over 60 representative LMMs demonstrate the effectiveness and reliability of both the pipeline and derived benchmarks, and show that LMMs performing well on existing benchmarks can still struggle with basic image understanding and detailed description, indicating room to improve these capabilities in future models. MMGenBench-Pipeline further enables efficient, annotation-free assessment of LMMs across diverse domains from raw image collections. All code and data will be released.

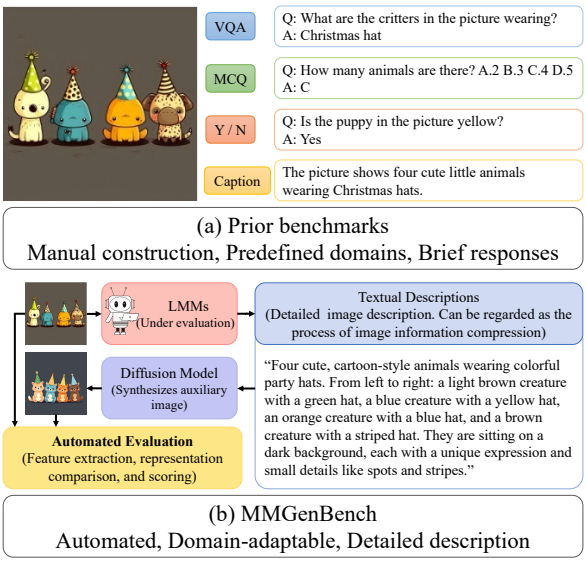


Figure 1: Comparison between prior benchmarks and MMGenBench. MMGenBench differs in three aspects: (1) it uses text-to-image and image representation models to automate LMM evaluation without manual annotation; (2) it supports domain-specific evaluation from raw image collections, whereas many prior benchmarks are tied to predefined domains; and (3) it evaluates detailed image descriptions rather than relying primarily on brief reference answers or closed-form responses.

1 Introduction

We have witnessed rapid progress in Large Multimodal Models (LMMs) (Huang et al., 2023; Wang et al., 2023; Lu et al., 2024b; Qwen Team, 2025b; Wang et al., 2025; OpenAI, 2025; Anthropic, 2026), which extend the strengths of LLMs (OpenAI, 2023; Team et al., 2023; OpenAI, 2024; Meta, 2024) to visual-textual processing. Compared with text, images are characterized by high abstraction and information density, while exhibiting strong spatial correlations and structural complexity. Reliable benchmarks are crucial for accurately evaluating LMMs. Many widely used benchmarks (Liu et al., 2025; Chen et al., 2024b; Yue et al., 2024a; Fu et al., 2023, 2024c; Liu et al., 2024b; Chen et al., 2024a; Ouyang et al., 2024) provide standardized evaluations across various multimodal tasks, yet they frequently depend on traditional datasets and are susceptible to data leakage and limited task diversity. Although tasks such as VQA (Lu et al., 2024a; Yu et al., 2023a; Ma et al., 2024; Liu et al., 2024c; Wang et al., 2024a; He et al., 2024), image captioning (Ke et al., 2019), and OCR (Liu et al., 2024b; Fu et al., 2024b; Yang et al., 2024) are care-

fully designed, as shown in Fig. 1(a), many existing benchmarks require costly manual construction and focus on specific domains, making them difficult to extend to new domains. Their reference answers also remain brief, primarily evaluating image understanding while overlooking the ability of LMMs to generate detailed visual descriptions.

In addition, image understanding and image generation largely remain separate research paradigms, with powerful models (Labs, 2024; Google DeepMind, 2025; OpenAI, 2026; Anthropic, 2026) following distinct modeling principles. For instance, GPT-5 (OpenAI, 2025), grounded in the next-token-prediction paradigm, exhibits impressive image comprehension, whereas Flux (Labs, 2024) and Seedream 4 (ByteDance Seed, 2025) achieve noteworthy success in text-to-image synthesis via diffusion models (Ho et al., 2020; Rombach et al., 2022). Furthermore, LMMs are extensively employed to generate data for generative models (Pei et al., 2024; Wu et al., 2024b; Lian et al., 2024; Huang et al., 2024a). This complementarity suggests that a sufficiently detailed understanding of an image should allow an LMM to distill its essential information into a text prompt from which a text-to-image model can reconstruct the salient visual content—a form of "compression". Hence, evaluating LMMs through text-to-image models is a principled and informative direction.

In this paper, we propose MMGenBench-Pipeline, an automated evaluation pipeline (refer to Fig. 1(b)) that prompts LMMs to generate textual descriptions from input images, employs text-to-image generative models to synthesize auxiliary images, and then uses an image representation model to obtain embeddings and compute post-processed scores for image understanding and description. To verify the effectiveness, we first introduce MMGenBench-Domain, which demonstrates how the pipeline can conduct domain-specific evaluation of LMMs by solely utilizing image inputs, and further build MMGenBench-Test, a benchmark probing the fundamental image understanding and description capabilities of LMMs across 13 distinct image patterns (refer to Fig. 3 and Appendix B.5). Extensive experiments on over 60 representative LMMs demonstrate the effectiveness and reliability of both the pipeline and the benchmark. Notably, numerous LMMs excelling on existing benchmarks can still struggle with basic image understanding and detailed description, highlighting substantial room for improvement. MMGenBench-Pipeline further enables efficient, annotation-free evaluation of LMMs across diverse domains from raw image collections, providing a flexible and scalable benchmarking tool.

In summary, our contributions are as follows:

- **Automated Evaluation Pipeline**: We propose MMGenBench-Pipeline, an automated pipeline that evaluates LMMs in image understanding and description by solely utilizing image inputs. By leveraging text-to-image and image representation models, it substantially reduces human involvement and improves the efficiency and objectivity of evaluation.
- **Comprehensive Benchmarks**: We construct MMGenBench-Domain, which assesses LMMs in the generated-image domain and showcases the pipeline's applicability to domain-specific evaluation from raw image collections, together with MMGenBench-Test, a benchmark probing basic image understanding and description across 13 image patterns.
- **Extensive Evaluation**: We evaluate over 60 representative LMMs, providing insights into their capabilities and limitations in basic image understanding and detailed description.

Please refer to Appendix A for related work.

2 The MMGenBench-Pipeline

2.1 Automatic Evaluation Pipeline

Built upon text-to-image generative models and image representation models, the proposed pipeline for evaluating LMMs consists of four sequential components: input construction, textual description generation, auxiliary image generation, and quantitative metric computation.

Input Construction. Our method supports domain-specific evaluation of LMM image understanding and description without requiring additional annotations. As shown in Stage 1 of Fig. 2, given an image from the target domain, we construct the multimodal input for the LMM under evaluation using predefined prompts and feed it into the subsequent stages.

Textual Description Generation. The input to the LMM consists of two parts: the image to be understood and the prompt that guides inference during evaluation. To standardize this stage, we use a manually crafted prompt $\mathbf{P}_{art}$, which specifies five components: role, definition explanation,

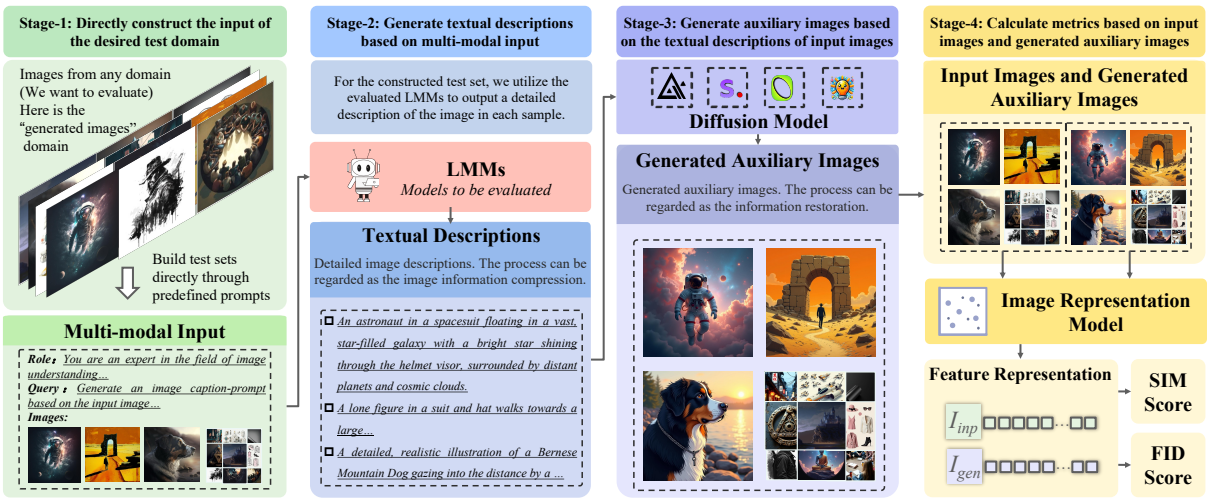


Figure 2: An overview of the MMGenBench-Pipeline, illustrating the automatic evaluation process. It starts by receiving user input (including the task instruction prompt and input images), and then generates the textual descriptions of input images. Subsequently, this process is followed by using a powerful text-to-image model to generate auxiliary images, then produces the representation of the input images and the generated ones using an image representation model, and finally outputs the evaluation score of LMMs.

task instruction, key points and requirements, and output format (see Appendix C.1 for details). This process is formalized as:

$$\mathbf{P}_{gen} = \text{LMM}(\mathbf{I}_{inp}, \mathbf{P}_{art}). \quad (1)$$

This yields a detailed textual description $\mathbf{P}_{gen}$ generated by the LMM, which serves as the input for the next stage.

Auxiliary Image Generation. In this stage, we use text-to-image models to generate auxiliary images from the given textual descriptions. Since generation quality depends on the chosen text-to-image model, we reduce generator-specific bias by selecting four representative models: FLUX.1-dev (Labs, 2024), Stable Diffusion 3.5 (Esser et al., 2024), Kolors (Team, 2024), and Lumina (Gao et al., 2024a). Comparing their results enables sensitivity analysis across generator backbones. This phase is represented as:

$$\mathbf{I}_{gen} = G(\epsilon; \mathbf{P}_{gen}, \theta), \quad (2)$$

where $\mathbf{I}_{gen}$ denotes the generated image, ϵ is a randomly sampled latent variable (typically Gaussian), θ denotes the model parameters, and G denotes the image generation function.

Quantitative Metric Computation. We quantify LMM performance by measuring the similarity between the input image $\mathbf{I}_{inp}$ and the generated image $\mathbf{I}_{gen}$. Since text-to-image models inevitably introduce stochasticity, pixel-level consistency is unattainable; we therefore compare images at the representation level. Specifically, for this stage, we adopt the Unicom (An et al., 2023) model as the default image encoder, and additionally consider DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021) as alternative encoders to assess the sensitivity of our metric across representation spaces (see Tables 18–21 for the sensitivity study). This phase is encapsulated as:

$$\mathbf{F}_{I_{inp}} = \text{Encoder}(\mathbf{I}_{inp}), \mathbf{F}_{I_{gen}} = \text{Encoder}(\mathbf{I}_{gen}), \quad (3)$$

where $\mathbf{F}_{I_{inp}}$ and $\mathbf{F}_{I_{gen}}$ denote the features of the input image $\mathbf{I}_{inp}$ and the generated image $\mathbf{I}_{gen}$, respectively. Based on these representations, we compute both a similarity score and a distribution-level generation quality score for quantitative evaluation.

2.2 Evaluation Metric

SIM-Score measures the similarity between two features, typically via cosine similarity. In our setting, it is computed as:

$$\text{SIM-Score}(\mathbf{I}_{inp}, \mathbf{I}_{gen}) = \frac{\mathbf{F}_{I_{inp}} \cdot \mathbf{F}_{I_{gen}}}{\|\mathbf{F}_{I_{inp}}\| \|\mathbf{F}_{I_{gen}}\|}, \quad (4)$$

where $\cdot$ denotes the dot product and $\|\cdot\|$ the vector norm used for normalization. SIM-Score ranges in $[-1, 1]$, with 1 indicating maximum similarity, 0 no similarity, and -1 complete opposition.

The FID-Score measures the discrepancy between the feature distributions of generated and reference images; a lower FID-Score indicates higher

generation quality and stronger distributional alignment with the reference images in the feature space. It is computed as:

$$\text{FID} = \|\mu_x - \mu_y\|^2 + \text{Tr}(\Sigma_x + \Sigma_y - 2(\Sigma_x \Sigma_y)^{1/2}), \quad (5)$$

where μ_x and μ_y denote the means of the input-image features $F_{I_{inp}}$ and generated-image features $F_{I_{gen}}$, and Σ_x, Σ_y denote their covariances. $\text{Tr}(\cdot)$ denotes the trace of a matrix.

3 The MMGenBench Benchmark

Since MMGenBench-Pipeline is automated and relies solely on raw image collections, it can be applied to evaluate LMMs in target domains without additional annotations: given a set of in-domain images, both SIM-Score and FID-Score can be computed directly. This property makes the pipeline a scalable and reproducible alternative to costly manual annotation, as the evaluation protocol transfers across domains whenever raw images are available. Leveraging this property, we construct **MMGenBench-Domain** by randomly sampling 10,000 images from the JourneyDB (Sun et al., 2024) validation set, which serves to assess LMMs on understanding and describing images from the generated-image domain. To further demonstrate that this design generalizes beyond the synthetic, style-rich JourneyDB distribution, we additionally evaluate MMGenBench on a real-photo subset constructed from COCO; details and results are reported in Sec. 4.5.

To further probe the fundamental ability of LMMs to perceive images and describe them in detail (as discussed in Sec. 1), we additionally introduce **MMGenBench-Test**, a curated test set comprising 1,284 images that span 13 image patterns drawn from the JourneyDB test set. These patterns are obtained through a multi-stage pipeline: GPT-4o first extracts coarse-grained candidate patterns, GPT-4-Turbo then consolidates them into a compact taxonomy of 13 patterns, and GPT-4o re-annotates every image under this fixed schema, with the final labels confirmed by human verification. The per-pattern distribution is reported in Fig. 3, where each pattern contains at least 114 images, and a single image may exhibit multiple patterns simultaneously. The complete construction pipeline, prompts, and per-pattern definitions are provided in Appendix B.

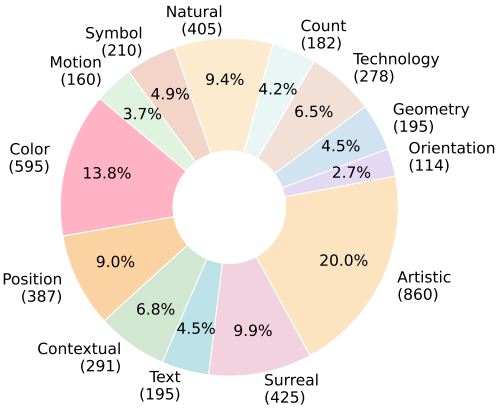


Figure 3: Statistics of MMGenBench-Test, which contains 1,284 images across 13 image patterns, with per-pattern sample counts shown. See Sec. 3 for details.

4 Experiment

4.1 Experimental Setup

Evaluation Models. For LMM evaluation, we selected over 60 models (refer to Appendix C.2) with strong performance on the VLM leaderboard (Duan et al., 2024). Our evaluation primarily covers open-source models, including Qwen2-VL (Wang et al., 2024b), Qwen3-VL (Qwen Team, 2025b), InternVL2 (Chen et al., 2023), InternVL3.5 (Wang et al., 2025), LLaVA-OV (Li et al., 2024b), Ovis (Lu et al., 2024b), among others. Additionally, we evaluated a broad set of frontier closed-source API models, including Claude Opus 4.6 (Anthropic, 2026), the GPT series (GPT-4o (OpenAI, 2024), GPT-5.2 and GPT-5.4 (OpenAI, 2025)), and the Qwen-VL API series (Qwen-VL-Max/Plus (Bai et al., 2023), Qwen2.5-VL-Max/Plus (Qwen Team, 2025a), Qwen3-VL-Plus/Flash (Qwen Team, 2025b)). For auxiliary image synthesis, we selected four text-to-image generation models: FLUX.1-dev (Labs, 2024), Stable Diffusion 3.5 (Esser et al., 2024), Kolors (Team, 2024) and Lumina (Gao et al., 2024a). During the final evaluation phase, we considered three widely used image representation models for feature extraction: Unicom (An et al., 2023), DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021). Unless otherwise specified, the reported results use Unicom as the default backbone, while results from DINOv2 and CLIP are provided in Sec. C.5 to demonstrate that our evaluation is stable across representation models.

Implementation Details. For textual description generation, we use the default settings of VLMEvalKit (Duan et al., 2024), modifying only

the predefined query prompt and the image to meet the task-specific requirements. Subsequently, we employ four distinct text-to-image models to generate auxiliary images and extract features using an image representation model. We then compute both the SIM-Score and FID-Score for evaluation under the same protocol. We analyze the performance metrics of four text-to-image models on the MMGenBench-Test dataset in Sec. C.4. The results indicate that the four text-to-image models show consistent trends in both SIM-Score and FID-Score. Therefore, unless otherwise stated, the reported results are obtained using FLUX.1-dev. Additional details are provided in Appendix C.2.

4.2 Main Results

Overall Performance. Table 1 reports results on both MMGenBench-Domain and MMGenBench-Test. MMGenBench-Domain covers the broad generated-image distribution, and its 10,000 images support more stable FID-Score estimation, while MMGenBench-Test's 13 patterns enable finer-grained diagnosis but make its FID-Score comparatively less stable. SIM-Score rankings on the two are highly consistent; therefore, we adopt SIM-Score as the primary metric and discuss MMGenBench-Test scores below. Even state-of-the-art LMMs are far from saturating MMGenBench-Test. Among closed-source models, Claude Opus 4.6 (Anthropic, 2026) ranks first with a SIM of 0.663, followed by the Qwen-VL API series (Qwen Team, 2025b) and the GPT series (OpenAI, 2025), whose best variants reach 0.657 and 0.637, respectively. The best open-source model, Qwen3-VL-32B (Qwen Team, 2025b), also reaches 0.637, matching GPT-5.4 and remaining close to the top closed-source results. Notably, there is no clear monotonic relationship between model size and performance across different model series, suggesting the importance of training data and protocols in developing LMMs with strong image understanding and detailed-description capabilities. Moreover, strong performance on existing benchmarks does not necessarily translate to strong performance on MMGenBench-Test; for example, LLaVA-OV (Li et al., 2024b) obtains a SIM-Score of 0.494. We also observe that newer series exhibit substantial gains over their predecessors: the open-source Qwen3-VL surpasses Qwen2-VL by a clear margin on the largest variant (0.637 vs. 0.553), the Qwen-VL API series steadily improves from Qwen-VL-Max through Qwen2.5-VL-Max to Qwen3-VL-Plus (0.552 → 0.604 → 0.657), and the GPT series improves from GPT-4o to GPT-5.4 (0.566 → 0.637); InternVL3.5 (Wang et al., 2025) also maintains competitive results against InternVL2 at comparable scales. Detailed results and analyses are provided in the following sections as well as in Appendix C.3-C.8.

Model	Test		Domain	
	SIM↑	FID↓	SIM↑	FID↓
Claude Opus 4.6	**0.663**	**1.039**	-	-
GPT-5.4	0.637	1.112	-	-
GPT-5.2	0.629	1.137	-	-
GPT-4o	0.566	1.306	-	-
Qwen3-VL-Plus	0.657	1.068	-	-
Qwen3-VL-Flash	0.651	1.093	-	-
Qwen2.5-VL-Max	0.604	1.232	-	-
Qwen2.5-VL-Plus	0.588	1.263	-	-
Qwen2-VL-Max	0.552	1.363	-	-
Qwen2-VL-Plus	0.475	1.586	-	-
Qwen3-VL-32B	0.637	1.149	**0.635**	**0.573**
Qwen3-VL-30B-A3B	0.608	1.219	0.607	0.616
Qwen3-VL-8B	0.610	1.196	0.608	0.617
Qwen2-VL-72B	0.553	1.357	0.545	0.710
Qwen2-VL-7B	0.532	1.437	0.524	0.775
Qwen2-VL-2B	0.487	1.549	0.501	0.806
InternVL3.5-38B	0.584	1.303	0.586	0.652
InternVL3.5-30B-A3B	0.573	1.317	0.572	0.677
InternVL3.5-8B	0.563	1.345	0.566	0.684
InternVL2-76B	0.599	1.264	0.599	0.632
InternVL2-40B	0.566	1.350	0.566	0.696
InternVL2-26B	0.576	1.320	0.577	0.671
InternVL2-8B	0.547	1.403	0.548	0.701
InternVL2-4B	0.556	1.345	0.556	0.689
InternVL2-2B	0.476	1.563	0.483	0.848
Ovis1.6-Gemma2-9B	0.582	1.316	0.579	0.667
Ovis1.5-Gemma2-9B	0.521	1.487	0.524	0.808
Ovis1.5-Llama3-8B	0.526	1.466	0.527	0.795
LLaVA-OV-72B	0.494	1.561	0.491	0.872
LLaVA-OV-SI-72B	0.512	1.512	0.514	0.813
LLaVA-OV-7B	0.488	1.590	0.490	0.861
LLaVA-OV-SI-7B	0.492	1.554	0.497	0.834
MiniCPM-V2.6-8B	0.548	1.386	0.545	0.710
RBDash-72B	0.525	1.413	0.527	0.740
xGen-MM-4.4B	0.414	1.671	0.412	0.940
MiniCPM-V2.5-8B	0.526	1.432	0.530	0.767

Table 1: Main results of different LMMs on MMGenBench-Domain and MMGenBench-Test, where SIM corresponds to SIM-Score, FID corresponds to FID-Score. More details are in Sec. C.3.

Pipeline Effectiveness. To validate the pipeline, we conducted a user study. We randomly selected 1850 samples from MMGenBench-Test. Each sample contained an original image and two outputs from different LMMs, which were randomly selected from 6 mainstream LMMs. Human annotators determined which LMM produced the better output, or whether the two outputs were compara-

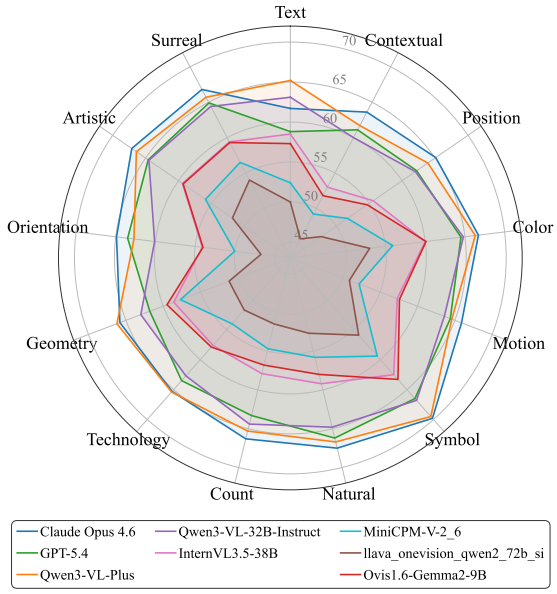


Figure 4: Model performance for representative evaluated models across image patterns on MMGenBench-Test. See Sec. 4.2 for detailed results and analysis.

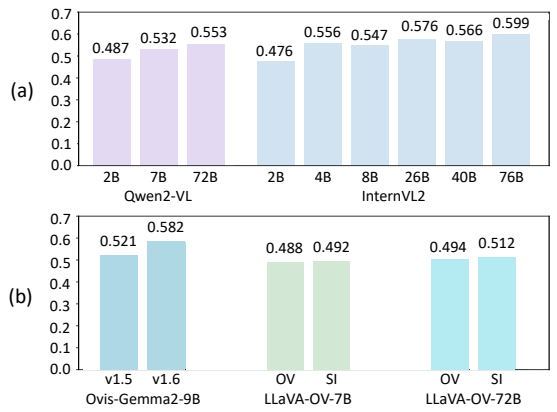


Figure 5: SIM-Score comparison of representative LMMs on MMGenBench-Test: (a) models from the same series with different parameter sizes; and (b) models with similar parameter sizes but different training protocols or data. See Sec. 4.2 for details.

ble. We then computed the SIM-Score. When the difference in SIM-Scores was less than 0.01, we treated the two outputs as comparable; otherwise, we selected the LMM with the higher score. We obtained the human metric from votes by 9 human experts. The results show that in 88.27% of cases, our pipeline aligns with human judgment. More details are provided in Appendix C.7.

Image Patterns Revealing Model's Strengths and Limitations. Fig. 4 shows the SIM-Score of the best-performing LMMs on the MMGenBench dataset. We observe that LMMs perform better on image patterns such as "Symbol", "Surreal", "Artistic", "Natural" and "Color". Conversely, their performance decreases on patterns such as "Contextual", "Orientation" and "Motion". This discrepancy suggests that LMMs are more proficient in coarse-grained image understanding and description. In contrast, fine-grained understanding and description remain more challenging, as these tasks require modeling complex contextual relationships within images and producing precise linguistic descriptions. Interestingly, on the latest closed-source frontier models (Claude Opus 4.6 and GPT-5.4), "Text" emerges as a relatively weak pattern, even when their overall scores rank highly: it becomes the lowest pattern in their own profiles (e.g., Claude Opus 4.6 reaches 0.698 on "Symbol" but only 0.617 on "Text"; GPT-5.4 drops from 0.602 on GPT-5.2 to 0.588). In contrast, Qwen3-VL-series models keep "Text" close to their overall mean. This indicates that recent style/safety-oriented training may bias frontier LMMs toward semantic abstraction rather than literal transcription of in-image text, which our image-to-text-to-image pipeline reveals through the resulting drop in visual reconstruction similarity.

Impact of Parameter Scales, Training Methods, and Data Variations. Fig. 5(a) shows the performance of variants from the same series with different parameter sizes, while Fig. 5(b) presents results for models with comparable parameter scales but different training protocols or datasets. As model size increases, Qwen2-VL improves from 0.487 to 0.553, and InternVL2 improves from 0.476 to 0.599. Additionally, with improved training data and protocols, Ovis1.6 outperforms Ovis1.5 by 0.06. Given that our evaluation is based on a single input image, LLaVA-OV-SI demonstrates stronger performance than LLaVA-OV.

4.3 Analysis

We identify three common problems that substantially affect the performance of LMMs in image understanding and description. Please refer to Appendix D for more details.

Failure in Following Instructions. As shown in Fig. 6, we select representative LMMs to illustrate a subset of qualitative examples. Certain LMMs do not strictly follow instructions when generating textual descriptions. Specifically, some LMMs prepend a prefix to the textual descriptions (e.g., RBDash-72B (RBDash-Team, 2024) in Fig. 6), while others append explanatory content after the textual descriptions (e.g., MiniCPM-Llama3-

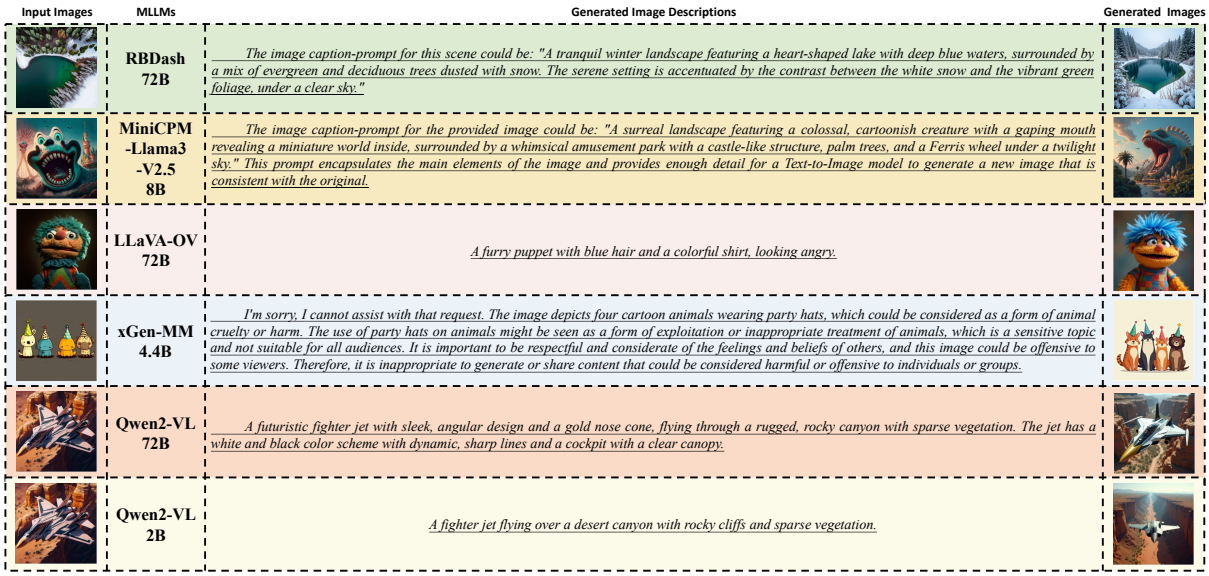


Figure 6: Qualitative examples on MMGenBench. The examples illustrate representative failure modes related to output format, generated content, and evaluation results. See Sec. 4.3 and Appendix D for additional discussion.

V2.5 (Yao et al., 2024) in Fig. 6). Additionally, instruction following does not appear to correlate consistently with parameter size across different LMM series. For example, both Qwen2-VL-2B and InternVL2-2B generate textual descriptions that follow the given instructions. We argue that an effective LMM, especially after instruction tuning, should be capable of following instructions and completing the task accurately.

Inability in Generating Detailed Textual Descriptions. Most image-text pairs used to train existing LMMs contain relatively short image captions. Consequently, numerous models exhibit suboptimal performance on the MMGenBench Benchmark, which requires the model to understand the image and provide a detailed description. As shown in Fig. 6, LLaVA-OV-72B (Li et al., 2024b), despite its 72B parameters, produces relatively short textual descriptions, which limits its performance on MMGenBench. Additionally, we observe that for LMMs in the same series, smaller LMMs tend to generate shorter textual descriptions, as evidenced by the Qwen2-VL and InternVL2 series. Based on these findings, we hypothesize that incorporating detailed image-description pairs during LMM training could improve performance.

Model Overfitting. Existing LMMs frequently undergo task-specific training to perform well on established benchmarks. This specialized training can lead to overfitting, which may compromise the basic image understanding and description capabilities of LMMs. For instance, the xGen-MM (Xue et al., 2024) results in Fig. 6 suggest overfitting toward the notion of "safety", despite the given image not presenting any safety risk. This observation suggests that when adapting models to various tasks, it is crucial to preserve basic image understanding and description capabilities, especially under distribution shifts.

4.4 Sensitivity to Pipeline Components

We assess the sensitivity of MMGenBench to pipeline components by sweeping four T2I generators (FLUX.1-dev, SD3.5, Kolors, Lumina) and three image encoders (Unicom (An et al., 2023), DINOv2 (Oquab et al., 2023), CLIP (Radford et al., 2021)). SIM-based LMM rankings correlate strongly across T2I generators (Spearman $= 0.98$, Kendall $= 0.90$) and across encoders (Spearman $= 0.98$, Kendall $= 0.90$); FID rankings are slightly more variable but remain consistent (Spearman ≥ 0.91). Moreover, the average per-model SIM standard deviation across T2I generators is only 0.013 (range 0.03), indicating that residual ranking shifts mainly arise from the instability of weaker LMMs rather than the generator choice. Full per-T2I and per-encoder breakdowns are provided in Sec. C.4 and C.5 in Appendix.

4.5 Real-World Domain Image Evaluation

To evaluate the generality of MMGenBench, we constructed a real-photo subset using 2,000 COCO images and evaluated 4 representative LMMs; the results are shown in Table 3.

Granularity	Metric	Correct / Total	Agreement	Discordant (CLIP / Ours)	McNemar p
Sample-level	CLIP image-caption similarity	607 / 1000	60.70%	63 / 356	$<1e{-}10$
	MMGenBench (Ours)	900 / 1000	**90.00%**		
Vote-level	CLIP image-caption similarity	5495 / 9000	61.06%	660 / 3109	$<1e{-}10$
	MMGenBench (Ours)	7944 / 9000	**88.27%**		

Table 2: Comparison with CLIP image-caption similarity on 1,000 human-labeled pairwise caption preferences. "Discordant (CLIP / Ours)" reports the McNemar contingency: pairs correct only by CLIP vs. only by MMGenBench.

Model	Test		Real-World	
	SIM↑	FID↓	SIM↑	FID↓
InternVL2-8B	0.5478	1.4037	0.4925	1.4353
Ovis1.6-Gemma2-9B	0.5825	1.3167	0.5441	1.3189
Qwen2-VL-7B	0.5323	1.4370	0.4984	1.3946
LLaVA-OV-7B	0.4888	1.5904	0.4344	1.7266

Table 3: Performance comparison on MMGenBench-Test and the real-world domain subset.

Model	$SIM_{<20}$	$SIM_{20\text{-}60}$	$SIM_{>60}$
InternVL2-8B	0.4968	0.5478	0.5701
Ovis1.6-Gemma2-9B	**0.5262**	**0.5825**	**0.5932**
Qwen2-VL-7B-Instruct	0.4911	0.5323	0.5602
LLaVA-OV-7B	0.4494	0.4888	0.5245

Table 4: SIM-Score under three caption-length intervals (<20, 20–60, and >60 words). The relative ranking of models remains unchanged across all three intervals.

We further conducted a human alignment check by randomly sampling 1,000 comparisons, each labeled by three annotators, and obtained a 92.1% (2763/3000) agreement with human judgment. These results indicate that MMGenBench remains effective on real-world photographic data, supporting the generality of our pipeline beyond synthetic and style-rich datasets.

4.6 Comparison with Caption-Image Similarity Metrics

We further compare MMGenBench against CLIP image-caption similarity, a representative caption-image similarity metric, on 1,000 human-labeled pairwise caption preferences. The protocol follows Sec. 4.2: each sample contains an original image and two LMM-generated captions, with 9 annotators per sample yielding both sample-level and vote-level human labels.

As shown in Table 2, MMGenBench reaches 90.00% sample-level and 88.27% vote-level agreement with humans, while CLIP image-caption similarity attains 60.70% and 61.06%, respectively. McNemar's exact test on paired predictions further confirms the gap: the discordant pairs are heavily skewed toward MMGenBench (356 vs. 63 at the sample level; 3109 vs. 660 at the vote level), with $p < 1e{-}10$ at both granularities. These results show that MMGenBench is substantially more aligned with human judgment than caption-image similarity metrics.

4.7 Length Bias of LMMs

A controlled study over four representative LMMs with three explicit word-count constraints (<20, 20–60, >60 words) shows that, although SIM/FID improve monotonically with caption length, the relative ordering of models is preserved across all three length intervals (Table 4). The benchmark therefore primarily reflects semantic content rather than verbosity, and the 20–60 word range adopted by default provides a stable and comparable evaluation regime. Per-model word-count statistics and full SIM/FID tables are reported in Appendix C.6.

5 Conclusion

In this paper, we presented MMGenBench-Pipeline as a generation-based evaluation framework for assessing whether LMMs can convert visual inputs into descriptions that preserve salient content. The pipeline uses text-to-image synthesis and representation level comparison to turn raw image collections into scalable evaluation protocols without task specific manual annotation. On this basis, MMGenBench-Domain evaluates generated-image understanding, while MMGenBench-Test examines basic image understanding and detailed description across 13 distinct image patterns. Results on over 60 representative LMMs show that the pipeline and benchmarks provide stable, informative signals, while also revealing remaining weaknesses in fine-grained perception and description even for models that score well on existing benchmarks. These findings suggest that generative reconstruction offers a complementary lens for reliably tracking progress in visual understanding and description quality. All code and data will be made publicly available.

Limitations and Future Work

MMGenBench-Pipeline relies on text-to-image and image representation models. Although the pipeline is designed to reduce evaluation noise, its accuracy may be limited for inputs that remain challenging for current text-to-image or image representation models, such as table images. This limitation suggests a clear direction for future work: target-task-aware rendering and matching pipelines may mitigate such failure modes; for example, in the context of table images, LaTeX could be used to render structured content and support pixel-level image matching.

In addition, introducing a text-to-image model and an image representation model increases evaluation time. Even so, MMGenBench-Pipeline remains substantially more cost-effective than manually constructed benchmarks, and its annotation-free protocol enables scalable evaluation of LMM performance across diverse domains. This scalability allows the pipeline to incorporate stronger models and updated datasets as they become available. As text-to-image models continue to advance, the current computational overhead is expected to decrease. For instance, FLUX.1-schnell requires only four denoising steps to generate images, which can substantially accelerate the evaluation process.

Ethical Considerations

This paper presents MMGenBench, an automated benchmark and evaluation pipeline for large multimodal models from a text-to-image generation perspective. By reducing reliance on manual annotation and enabling scalable, reproducible assessment across diverse domains, our approach can help diagnose fine-grained description and instruction-following failures and support more reliable multimodal systems. As with other benchmark resources, MMGenBench should be interpreted as a diagnostic tool rather than a comprehensive measure of model safety or societal impact. Future uses should consider potential biases inherited from the underlying image collections, text-to-image generators, and image representation models.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, and 1 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Xiang An, Jiankang Deng, Kaicheng Yang, Jaiwei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. Unicom: Universal and compact representation learning for image retrieval. *Preprint*, arXiv:2304.05884.

Anthropic. 2026. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, and 1 others. 2024. Benchmarking foundation models with language-model-as-an-examiner. *NIPS*, 36.

Han Bao, Yue Huang, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuyin Chen, Mohamed Elhoseiny, and Xiangliang Zhang. 2024. Autobench-v: Can large vision-language models benchmark themselves? *arXiv preprint arXiv:2410.21259*.

ByteDance Seed. 2025. Seedream 4.0: Toward next-generation multimodal image generation. https://arxiv.org/abs/2509.20427.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, and 1 others. 2024a. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. corr, abs/2410.10563, 2024b. doi: 10.48550. *arXiv preprint ARXIV.2410.10563*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, and 1 others. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *Preprint*, arXiv:2407.11691.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, and 1 others. 2024b. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Preprint*, arXiv:2501.00321.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024c. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024d. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Chen Lin, Rongjie Huang, Shijie Geng, Renrui Zhang, Junlin Xi, and 1 others. 2024a. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *Preprint*, arXiv:2405.05945.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, and 1 others. 2024b. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*.

Team GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Google DeepMind. 2025. Gemini 2.5 flash image (nano banana). `https://deepmind.google/models/gemini/image/`.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *Preprint*, arXiv:2402.14008.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, and 1 others. 2024. Cogvlm2: Visual language models for image and video understanding. *Preprint*, arXiv:2408.16500.

Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, and 1 others. 2023. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29.

Kaiyi Huang, Yukun Huang, Xuefei Ning, Zinan Lin, Yu Wang, and Xihui Liu. 2024a. Genmac: Compositional text-to-video generation with multi-agent collaboration. *Preprint*, arXiv:2412.04440.

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. 2024b. Mini-monkey: Alleviating the semantic sawtooth effect for lightweight mllms via complementary image pyramid. *Preprint*, arXiv:2408.02034.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.

Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. 2019. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8888–8897.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024a. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024b. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. 2024a. Genai-bench: Evaluating and improving compositional text-to-visual generation. *Preprint*, arXiv:2406.13743.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. 2024c. Autobencher: Creating salient, novel, difficult datasets for language models. *arXiv preprint arXiv:2407.08351*.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2024. Llm-grounded video diffusion models. *Preprint*, arXiv:2309.17444.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pre-training for visual language models. *Preprint*, arXiv:2312.07533.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. 2024b. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Preprint*, arXiv:2305.07895.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024c. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *Preprint*, arXiv:2406.11833.

DataCanvas Ltd. 2024. Mmalaya2. https://huggingface.co/DataCanvas/MMAlaya2.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, and 1 others. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024b. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Preprint*, arXiv:2407.01523.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. In *Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/.

OpenAI. 2026. Gpt image 2. https://openai.com/index/introducing-chatgpt-images-2-0/.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2024. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *Preprint*, arXiv:2412.07626.

Yuhan Pei, Ruoyu Wang, Yongqi Yang, Ye Zhu, Olga Russakovsky, and Yu Wu. 2024. Sowing information: Cultivating contextual coherence with mllms in image generation. *Preprint*, arXiv:2411.19182.

Qwen Team. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Qwen Team. 2025b. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

RBDash-Team. 2024. Rbdash-v1.2-72b. https://huggingface.co/RBDash-Team/RBDash-v1.2-72b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, and 1 others. 2025. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *Preprint*, arXiv:2408.15998.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, and 1 others. 2024. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kolors Team. 2024. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Zero Vision. 2024. Llama-3-mixsensev1_1. https://huggingface.co/Zero-Vision/Llama-3-MixSenseV1_1.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *Preprint*, arXiv:2406.09411.

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. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482.

WeMM. 2024. Wemm. https://github.com/scenarios/WeMM.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations (ICLR)*.

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, and 1 others. 2024a. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*.

Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. 2024b. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336.

XinYuan. 2024. Xinyuan. https://huggingface.co/Cylingo/Xinyuan-VL-2B.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, and 1 others. 2024. xgen-mm (blip-3): A family of open large multimodal models. *Preprint*, arXiv:2408.08872.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. 2024. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *Preprint*, arXiv:2412.02210.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin

Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023a. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2409.02813.

Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Task me anything. *arXiv preprint arXiv:2406.11775*.

Tiancheng Zhao, Qianqian Zhang, Kyusong Lee, Peng Liu, Lu Zhang, Chunxin Fang, Jiajia Liao, Kelei Jiang, Yibo Ma, and Ruochen Xu. 2024. Omchat: A recipe to train multimodal language models with strong long context and video understanding. *Preprint*, arXiv:2407.04923.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024. Dynamic evaluation of large language models by meta probing agents. In *Forty-first International Conference on Machine Learning*.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jurgen Schmidhuber. 2024. Agent-as-a-judge: Evaluate agents with agents. *Preprint*, arXiv:2410.10934.

A Related Work

A.1 Automated Benchmarks

The rapid progress of LLMs has motivated a broad line of benchmark construction methods for scalable and reproducible evaluation. Efforts such as LMExamQA (Bai et al., 2024) introduced the "Language-Model-as-an-Examiner" framework to support scalable LLM assessment. Other work has addressed more specific limitations in reasoning-oriented evaluation, particularly in dynamic settings. DYVAL (Zhu et al., 2023) focuses on dynamic reasoning ability, whereas DYVAL2 (Zhu et al., 2024) extends this direction through meta probing agents inspired by psychometrics, enabling multifaceted analysis of language understanding, problem solving, and domain knowledge. AutoBencher (Li et al., 2024c) further automates the construction of challenging datasets for LLM evaluation. Related benchmark-generation frameworks also reduce manual construction costs: UNIGEN (Wu et al., 2024a) supports LLM-based synthetic text dataset generation and dynamic benchmarking, while Task Me Anything (Zhang et al., 2024) generates user-tailored multimodal benchmarks for perceptual capability evaluation. Recent studies have also targeted data contamination and human-grading bottlenecks: MixEval (Ni et al., 2024) constructs evaluation sets by matching web-mined user queries with similar examples from existing benchmarks, LiveBench (White et al., 2025) maintains frequently updated, contamination-limited questions that are automatically scored against objective ground truth, and Agent-as-a-Judge (Zhuge et al., 2024) extends the LLM-as-a-Judge paradigm to agentic systems, providing intermediate feedback across the task-solving process.

A.2 LMM Benchmarks

The emergence of LMMs has highlighted the limitations of task-isolated evaluations, motivating fine-grained and contamination-aware benchmarks (Liu et al., 2025; Chen et al., 2024b; Yu et al., 2023b). MMStar (Chen et al., 2024b) mitigates data leakage and visual-shortcut risks, while AUTOBENCH-V (Bao et al., 2024) automates the benchmarking process to improve scalability and objectivity. Recent efforts further move LMM evaluation toward fine-grained capability diagnosis: MMMU-Pro (Yue et al., 2024b) strengthens MMMU by removing text-solvable questions and introducing a vision-only setting; BLINK (Fu et al., 2024d) shows that core perception abilities solvable by humans "within a blink" remain challenging for state-of-the-art LMMs; Video-MME (Fu et al., 2024a) extends evaluation to long-form video understanding; and HallusionBench (Guan et al., 2024) diagnoses entangled language hallucination and visual illusion. Despite this progress, many widely used benchmarks still rely on curated multiple-choice or VQA-style questions, which couple LMM evaluation to question design and human-annotated answers, and provide limited coverage of open-ended image description quality. From the text-to-image evaluation side, GenAI-Bench (Li et al., 2024a) reports that VQA-style scoring aligns substantially better with human judgment than CLIP-based similarity for compositional text-to-visual generation. In contrast, our work uses a frozen text-to-image generator to probe LMM image understanding and description quality. This generative round-trip removes the dependence on hand-crafted questions and supports scalable, annotation-free evaluation as new LMMs emerge.

B Details of MMGenBench Benchmark

B.1 Overview

As emphasized in Sec. 1, MMGenBench-Pipeline is automated and relies solely on raw images, enabling domain-specific evaluation without additional annotations. Building on this property, we first construct MMGenBench-Domain, a domain-level dataset in the generated-image domain, to evaluate LMMs' ability to understand and describe generated images; it is built by sampling images from the JourneyDB (Sun et al., 2024) validation set with no additional annotation. To further probe basic image understanding and description capabilities across diverse image types, we additionally construct MMGenBench-Test, a curated test set covering 13 image patterns from the JourneyDB test set. We use a multi-stage procedure to extract and annotate image patterns, as illustrated in Fig. 8, and manually double-check the image patterns before finalizing the annotations. Beyond these two datasets, MMGenBench-Pipeline can be applied to other domains by replacing the input image collection with images from the target domain.

B.2 Dataset Statistics

MMGenBench-Domain is constructed by randomly sampling 10,000 images from the JourneyDB val-

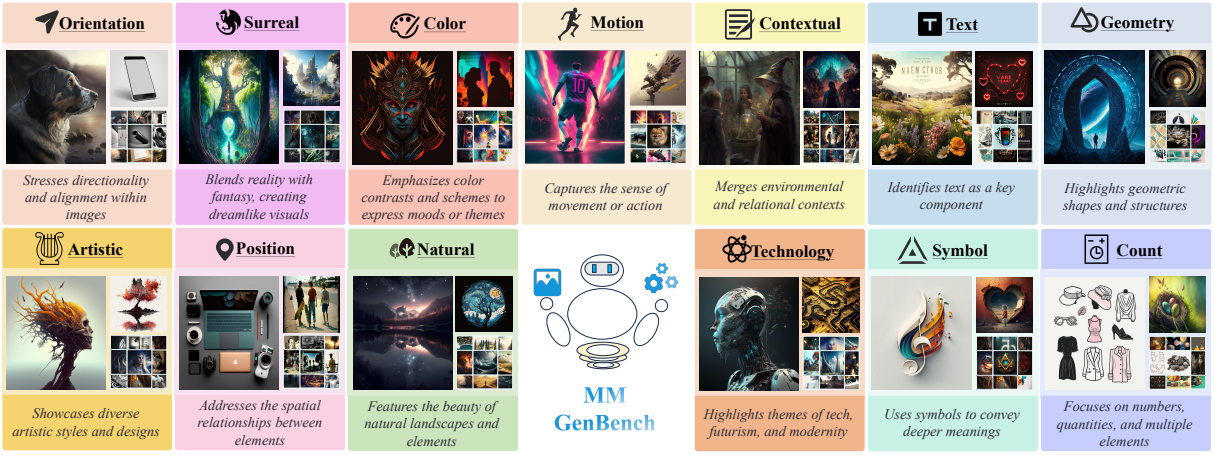


Figure 7: Examples from MMGenBench-Test. The dataset contains 13 distinct image patterns, and each pattern is illustrated with representative images and a concise textual explanation.

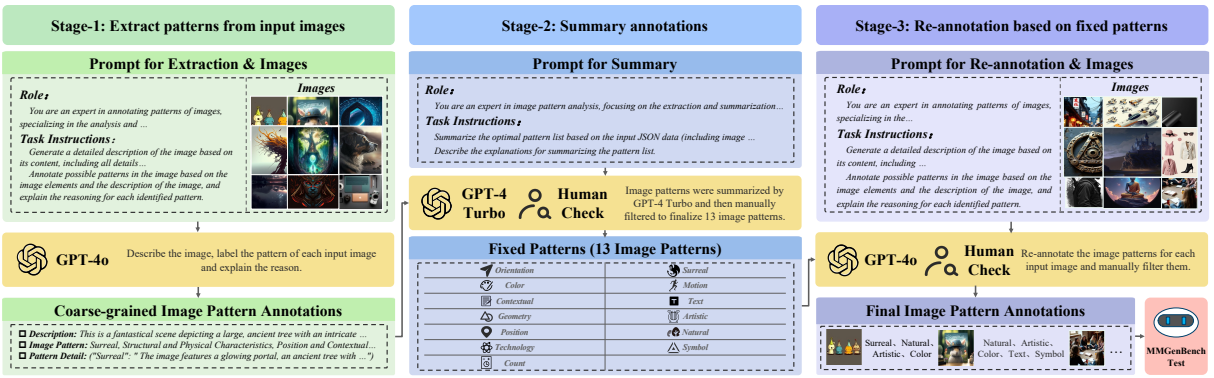


Figure 8: Overview of the MMGenBench-Test construction process. We use GPT-4o to extract image patterns from input images, GPT-4 Turbo to summarize the extracted patterns, and human verification to finalize 13 image patterns. GPT-4o is then used for re-annotation, followed by human review to produce the final annotations.

idation set, which allows us to evaluate the image understanding and description performance of LMMs within the generated-image domain without any additional annotation. MMGenBench-Test contains 1,284 images across 13 distinct image patterns (see Appendix B.5). The distribution of images per pattern is shown in Fig. 3, where each pattern contains at least 114 images. An image may contain multiple patterns; for instance, the first image annotation in Fig. 8 contains four image patterns: "Surreal", "Natural", "Artistic" and "Color". These 13 image patterns are designed to measure the image comprehension and description capabilities of LMMs across diverse dimensions, and all samples are manually checked and annotated to ensure overall quality.

B.3 MMGenBench-Domain Construction

Constructing high-quality datasets typically requires substantial annotation effort. Therefore, efficient methods for creating domain-specific evaluation sets are important when assessing the image understanding and description capabilities of LMMs in new domains. By leveraging MMGenBench-Pipeline, our method supports LMM evaluation across different domains as long as domain-specific images are available. In this study, we randomly select 10,000 images from the JourneyDB validation set to construct MMGenBench-Domain. To quantify LMM performance within this domain, we compute both FID-Score and SIM-Score.

B.4 MMGenBench-Test Data Collection

Pattern Extraction. We extract image patterns from existing images to measure the understanding and description ability of LMMs across various image categories. Because the construction setting assumes only image inputs are available, we use GPT-4o to extract and analyze underlying image patterns. The task is executed by GPT-4o in three sequential steps: 1) providing a detailed description

of the image; 2) annotating the possible patterns based on the image features and description; and 3) explaining the rationale behind the annotated image patterns. With the prompts described in Appendix B.7, GPT-4o produces image descriptions, pattern annotations, and rationales for the target images. An example is presented in Fig. 9 in Appendix B.6, where the image description, extracted image patterns, and rationales are consistent with the image content. By annotating each image in the JourneyDB test set, we obtain a total of 1868 image patterns.

Annotation Summary. By leveraging GPT-4o, we extract a broad set of image patterns and observe that some patterns overlap semantically. We compute frequency statistics for the extracted patterns and rank them in descending order (as shown in Appendix B.5). We then use GPT-4-Turbo to summarize these patterns with the prompt shown in Appendix B.7. This procedure results in a curated list of 13 distinct image patterns, each described in Appendix B.5.

Image Re-annotation. We re-annotate the JourneyDB test set using the previously identified 13 image patterns. Similar to the pattern extraction stage, GPT-4o is used to generate descriptions for each image. These descriptions are then annotated with the predefined image patterns, together with the corresponding rationales. A key difference in this stage is that re-annotation is restricted to the specified 13 image patterns and excludes other patterns. Given the size of the test set, we select a subset of images for subsequent processing. The re-annotated images are classified into distinct image patterns, each containing a specified number of images. To construct the final dataset, we randomly sample a subset of images from each pattern with a selection probability of $\frac{100}{N}$, where N represents the total number of images within the pattern. To mitigate potential annotation errors, we manually verify and annotate each image. The final dataset is established by voting between the model predictions and annotations.

B.5 Details of MMGenBench-Test Data

Thirteen Image Patterns. As illustrated in Fig. 7 and 10, summarization is performed using GPT-4 Turbo, followed by human verification. We identify 13 distinct image patterns, each with a detailed explanation. A single image may exhibit multiple patterns concurrently.

Patterns Extracted by GPT-4o. Fig. 11 shows the image patterns extracted by GPT-4o, organized in descending order. While the number of extracted patterns is 1868, only the most frequently occurring patterns are presented due to space constraints.

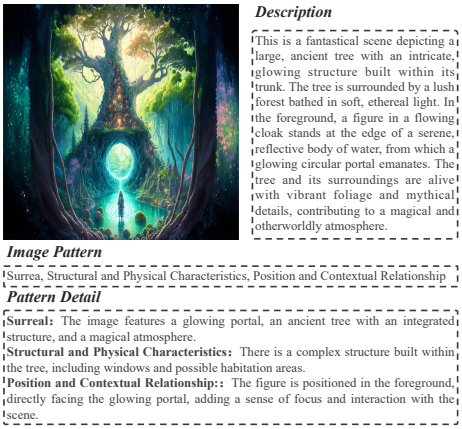


Figure 9: Case study of image pattern extraction. GPT-4o generates an image description, identifies underlying image patterns, and provides rationales for the extracted patterns.

B.6 Case Study of Pattern Extraction

Fig. 9 illustrates an example in which GPT-4o identifies image patterns from a given image. The results show clear correspondence among the image description, the extracted patterns, and the corresponding explanations, all of which are consistent with the visual content.

B.7 Details of Construction Prompts

To construct MMGenBench-Test, we design a series of prompts tailored to the requirements of each stage. Fig. 12, 13, and 14 correspond to the three sequential stages in the pipeline (see Fig. 2) used to develop MMGenBench-Test.

C Additional Experimental Details

C.1 Evaluation Pipeline Prompt

We design a prompt to guide LMMs in generating textual descriptions of images for subsequent evaluation. As shown in Fig. 15, the prompt covers five perspectives: “role”, “definition”, “task”, “key points”, and “output”. To maintain evaluation consistency across models, we use the same prompt for all evaluated LMMs.

C.2 Experimental Setup

We evaluate over 60 representative LMMs, as detailed in Tables 9 and 10, which are available on the OpenVLM Leaderboard (Contributors, 2023). For

open-source LMMs, we use the original configuration of VLMEvalKit (Duan et al., 2024) to generate textual descriptions of images on 8 NVIDIA-H20 GPUs. During the image generation and metric calculation stages, we also use 8 NVIDIA-H20 GPUs for multi-process inference.

C.3 Comprehensive Experimental Evaluation

A comprehensive set of experimental results covering over 60 representative LMMs is shown in Tables 9 and 10.

C.4 Sensitivity Analysis of T2I Models

We conducted a comprehensive sensitivity analysis on four different text-to-image models (FLUX.1-dev, SD3.5, Kolors, Lumina). The results show:
(1) **LMM rankings are stable across T2I models.** SIM correlations across the four generators are high (Spearman $= 0.98$, Kendall $= 0.90$ on average). This indicates that changing the T2I model has limited effect on the relative ranking of LMMs (Tables 12, 13).
(2) **FID is more sensitive, but still maintains strong cross-model consistency.** FID correlations remain high (Spearman $= 0.91$, Kendall $= 0.78$), consistent with its known dependence on generator style distribution (Tables 14, 15).
(3) **Ranking variation is more pronounced for weaker LMMs.** Only a small subset of weaker LMMs exhibits noticeable variance across T2I models, while strong LMMs remain stable. This suggests that the rankings are not dominated by artifacts of a specific T2I model (Table 11).
(4) **MMGenBench evaluations are consistent across generators.** The average variation of SIM-Scores across T2I models is small (std $= 0.013$, range $= 0.03$), suggesting that the benchmark is not dominated by T2I controllability (Table 17).

Overall, these results indicate that the scores are not tightly entangled with any particular text-to-image model, and the benchmark conclusions are consistent across different text-to-image models.

C.5 Sensitivity Analysis of Embedding Models

We conduct a sensitivity analysis across three widely used embedding models (Unicom/ViT-L-14-336px (An et al., 2023), DINOv2/dinov2-large (Oquab et al., 2023), CLIP/clip-vit-large-patch14-336 (Radford et al., 2021)). The results (Tables 18–21) show that MMGenBench is stable across embedding models. SIM-Scores exhibit high cross-model correlations (Spearman $= 0.98$, Kendall $= 0.90$), and FID-Scores are even more consistent (Spearman $= 0.99$, Kendall $= 0.92$). These correlations indicate that the LMM rankings are largely unchanged across Unicom, DINOv2, and CLIP, suggesting that the evaluation is not sensitive to the embedding backbone.

C.6 Length Bias of LMM

We conducted a controlled study on four representative LMMs (InternVL2-8B, Ovis1.6-Gemma2-9B, Qwen2-VL-7B-Instruct, LLaVA-OneVision-Qwen2-7B) under three explicit word-count constraints: < 20, $20 - 60$ (our default), and > 60 words (Table 5).

As shown in Table 6, across all models, SIM/FID improves monotonically with length (e.g., InternVL2-8B: SIM $0.4968 \rightarrow 0.5478 \rightarrow 0.5701$; Ovis1.6-Gemma2-9B: $0.5262 \rightarrow 0.5825 \rightarrow 0.5932$), indicating that very short captions (< 20 words) may omit essential semantics. However, the ordering of models remains unchanged across all three length intervals, and the gains for > 60 word outputs are modest. This suggests that the benchmark is not driven solely by verbosity, but also reflects the amount of semantic content conveyed. Based on these observations, the 20–60 word range provides a stable and comparable setting.

C.7 Human Alignment Score

We describe the experimental setup, annotator protocol, and statistical results for the human alignment study.

User Study Setup. We select 6 representative LMMs: GPT-4o-high, InternVL2-76B, Qwen2-VL-72B-Instruct, LLaVA-OneVision-Qwen2-72B, MiniCPM-V-2.6, and Ovis1.6-Gemma2-9B. We construct 1,000 paired caption comparisons, where each comparison is formed by: (1) randomly sampling two models from the six, (2) randomly selecting one image from MMGenBench-Test, and (3) presenting the two model-generated captions to human annotators.

Annotator Task. Each of the 1,000 examples is independently labeled by 9 human annotators. For each example, annotators are instructed to choose:

- **C1** if caption-1 describes the image better,
- **C2** if caption-2 describes the image better, or
- **B** if the two captions are comparable.

For example, an annotation list

Model	< 20 Mean	< 20 Std	20-60 Mean	20-60 Std	> 60 Mean	> 60 Std
InternVL2-8B	19.17	4.70	38.87	12.51	81.81	83.38
Ovis1.6-Gemma2-9B	22.49	6.55	55.85	12.26	90.43	18.39
Qwen2-VL-7B-Instruct	15.48	21.06	27.67	9.08	68.36	66.71
LLaVA-OneVision-Qwen2-7B	11.09	3.37	17.35	6.99	36.75	24.88

Table 5: Word-count statistics under different caption-length constraints.

Model	< 20 SIM	< 20 FID	20–60 SIM	20–60 FID	> 60 SIM	> 60 FID
InternVL2-8B	0.4968	1.5337	0.5478	1.4037	0.5701	1.3171
Ovis1.6-Gemma2-9B	0.5262	1.4545	0.5825	1.3167	0.5932	1.2645
Qwen2-VL-7B-Instruct	0.4911	1.5562	0.5323	1.4370	0.5602	1.3651
LLaVA-OneVision-Qwen2-7B	0.4494	1.7247	0.4888	1.5904	0.5245	1.4734

Table 6: SIM-Score and FID-Score under different caption-length constraints.

```
['C1','C1','C1','C1','C1','C1','C1','C2','B']
```

indicates that 7 annotators prefer caption-1, 1 prefers caption-2, and 1 considers them similar.

Agreement Computation. Our pipeline predicts the outcome using the SIM-score difference:

- $\Delta > 0.01 \Rightarrow$ **C1**,
- $\Delta < -0.01 \Rightarrow$ **C2**,
- $|\Delta| \leq 0.01 \Rightarrow$ **B**.

We compute agreement against all $9 \times 1{,}000 = 9{,}000$ individual human votes. The pipeline shows substantial alignment with human judgments:

- Agreement with individual human votes: $7944/9000 = 88.27\%$,
- Agreement with human majority decisions ($\geq 5/9$ votes): $900/1000 = 90.0\%$.

Pattern-Wise Agreement. To investigate whether certain patterns are inherently more ambiguous or less consistent, we further compute vote agreement separately for patterns with sufficiently large numbers of human-evaluated samples (approximately ≥ 100), which provides more reliable pattern-level estimates.

Pattern	Right	Wrong	Agreement (%)
Natural	216	23	90.4
Symbol	159	17	90.3
Technology	125	15	89.3
Artistic	111	16	87.4
Count	99	10	90.8
Motion	109	9	92.4

Table 7: Pattern-wise agreement between MMGenBench and human judgments.

Model	SIM↑		FID↓	
	Mean	Std	Mean	Std
GPT-4o	0.5656	0.0014	1.3110	0.0110
Qwen-VL-Max	0.5496	0.0033	1.3763	0.0047
Qwen-VL-Plus	0.4759	0.0024	1.5910	0.0115
Qwen2-VL-72B	0.5501	0.0009	1.3687	0.0060
Qwen2-VL-7B	0.5297	0.0014	1.4334	0.0085
Qwen2-VL-2B	0.4890	0.0016	1.5399	0.0069
InternVL2-76B	0.5985	0.0015	1.2622	0.0100
InternVL2-40B	0.5645	0.0008	1.3603	0.0059
InternVL2-26B	0.5782	0.0028	1.3147	0.0100
InternVL2-8B	0.5453	0.0002	1.3849	0.0022
InternVL2-4B	0.5584	0.0008	1.3402	0.0056
InternVL2-2B	0.4763	0.0038	1.5712	0.0132
Ovis1.6-Gemma2-9B	0.5836	0.0005	1.3010	0.0092
Ovis1.5-Gemma2-9B	0.5245	0.0011	1.4883	0.0098
Ovis1.5-Llama3-8B	0.5263	0.0009	1.4629	0.0116
LLaVA-OV-72B	0.4962	0.0010	1.5620	0.0093
LLaVA-OV-SI-72B	0.5124	0.0020	1.5094	0.0118
LLaVA-OV-7B	0.4866	0.0019	1.6048	0.0134
LLaVA-OV-SI-7B	0.4925	0.0022	1.5622	0.0097
MiniCPM-V2.6-8B	0.5458	0.0014	1.3786	0.0039
RBDash-72B	0.5239	0.0019	1.4277	0.0090
xGen-MM-4.4B	0.4107	0.0021	1.6727	0.0141
MiniCPM-V2.5-8B	0.5276	0.0028	1.4402	0.0130

Table 8: Mean and Std of SIM and FID over five random seeds on MMGenBench-Test.

Key Takeaway. Across all evaluated patterns, agreement between the proposed metric and human judgment ranges from 87% to 92%. These patterns cover a large portion of MMGenBench-Test and represent diverse visual attributes, including object fidelity (Natural), symbolic semantics (Symbol), dynamic information (Motion), and fine-grained detail (Count). Overall, the results indicate that the pipeline aligns consistently with human perception across diverse image types.

C.8 Seed Variance of Results

To quantify variance, we conduct a five-seed evaluation on MMGenBench-Test and report the mean and standard deviation of SIM and FID for 23 main

models. Across all models, the observed stochastic variation is small: the standard deviation of SIM ranges from 0.0002 to 0.0038, while that of FID ranges from 0.0022 to 0.0141. Moreover, the relative ranking of the models remains consistent across different seeds, with Spearman correlations of $\rho = 0.9975$ for SIM and $\rho = 0.9970$ for FID.

As shown in Table 8, these results indicate that multi-run fluctuations are limited, and the scores reported in Table 1 closely match the multi-seed averages, with a maximum difference of 0.004. Therefore, the evaluation pipeline is stable with respect to randomness in text-to-image sampling.

D Qualitative Example

Qualitative analyses for over 60 representative LMMs are illustrated in Fig. 16-20.

Model	Test		Domain	
	SIM↑	FID↓	SIM↑	FID↓
Claude Opus 4.6 (Anthropic, 2026)	0.663	1.039	-	-
Qwen3-VL-Plus (Qwen Team, 2025b)	0.657	1.068	-	-
Qwen3-VL-Flash (Qwen Team, 2025b)	0.651	1.093	-	-
Qwen3-VL-32B-Instruct (Qwen Team, 2025b)	0.637	1.149	0.635	0.573
GPT-5.4 (OpenAI, 2025)	0.637	1.112	-	-
GPT-5.2 (OpenAI, 2025)	0.629	1.137	-	-
Qwen3-VL-8B-Instruct (Qwen Team, 2025b)	0.610	1.196	0.608	0.617
Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025b)	0.608	1.219	0.607	0.616
Qwen2.5-VL-Max (Qwen Team, 2025a)	0.604	1.232	-	-
Qwen2.5-VL-Plus (Qwen Team, 2025a)	0.588	1.263	-	-
InternVL3.5-38B (Wang et al., 2025)	0.584	1.303	0.586	0.652
InternVL3.5-30B-A3B (Wang et al., 2025)	0.573	1.317	0.572	0.677
InternVL3.5-8B (Wang et al., 2025)	0.563	1.345	0.566	0.684
Qwen2-VL-72B-Instruct (Wang et al., 2024b)	0.553	1.357	0.545	0.710
qwen-vl-max-0809 (Bai et al., 2023)	0.552	1.363	-	-
InternVL2-76B (Chen et al., 2023)	0.599	1.264	0.599	0.632
GPT4o_HIGH (OpenAI, 2024)	0.566	1.306	-	-
InternVL2-40B (Chen et al., 2023)	0.566	1.350	0.566	0.696
Ovis1.6-Gemma2-9B (Lu et al., 2024b)	0.582	1.316	0.579	0.667
llava_onevision_qwen2_72b_ov (Li et al., 2024b)	0.494	1.561	0.491	0.872
llava_onevision_qwen2_72b_si (Li et al., 2024b)	0.512	1.512	0.514	0.813
Qwen2-VL-7B-Instruct (Wang et al., 2024b)	0.532	1.437	0.524	0.775
InternVL2-26B (Chen et al., 2023)	0.576	1.320	0.577	0.671
qwen-vl-plus (Bai et al., 2023)	0.475	1.586	-	-
MiniCPM-V-2_6 (Yao et al., 2024)	0.548	1.386	0.545	0.710
InternVL2-8B (Chen et al., 2023)	0.547	1.403	0.548	0.701
Ovis1.5-Gemma2-9B (Lu et al., 2024b)	0.521	1.487	0.524	0.808
RBDash_72b (RBDash-Team, 2024)	0.525	1.413	0.527	0.740
MMAlaya2 (Ltd., 2024)	0.537	1.423	0.541	0.747
Ovis1.5-Llama3-8B (Lu et al., 2024b)	0.526	1.466	0.527	0.795
OmChat (Zhao et al., 2024)	0.439	1.761	0.447	1.028
InternVL-Chat-V1-5 (Chen et al., 2024c)	0.547	1.393	0.547	0.729
llava_onevision_qwen2_7b_si (Li et al., 2024b)	0.492	1.554	0.497	0.834
XComposer2d5 (Chen et al., 2024c)	0.460	1.646	0.458	0.955
Pixtral-12B (Agrawal et al., 2024)	0.552	1.372	0.561	0.686
InternVL2-4B (Chen et al., 2023)	0.556	1.345	0.556	0.689
llava_onevision_qwen2_7b_ov (Li et al., 2024b)	0.488	1.590	0.490	0.861
glm-4v-9b (GLM, 2024)	0.511	1.463	0.527	0.730
molmo-7B-D-0924 (Deitke et al., 2024)	0.519	1.490	0.523	0.813
XComposer2_4KHD (Chen et al., 2024c)	0.433	1.781	0.435	1.075
MiniCPM-Llama3-V-2_5 (Yao et al., 2024)	0.526	1.432	0.530	0.767
VILA1.5-40b (Lin et al., 2023)	0.483	1.656	0.486	0.958
cambrian_34b (Tong et al., 2024)	0.546	1.377	0.549	0.701
WeMM (WeMM, 2024)	0.488	1.565	0.479	0.888
Llama-3.2-11B-Vision-Instruct (Meta, 2024)	0.526	1.428	0.525	0.768
Qwen2-VL-2B-Instruct (Wang et al., 2024b)	0.487	1.549	0.501	0.806
XComposer2 (Chen et al., 2023)	0.248	2.890	0.247	2.130
cogvlm2-llama3-chat-19B (Hong et al., 2024)	0.465	1.523	0.464	0.825
Idefics3-8B-Llama3 (Laurençon et al., 2024a)	0.460	1.718	0.458	1.004
Mini-InternVL-Chat-4B-V1-5 (Gao et al., 2024b)	0.527	1.433	0.525	0.765
XinYuan-VL-2B-Instruct (XinYuan, 2024)	0.395	1.951	0.392	1.240
llava_next_yi_34b (Liu et al., 2024a)	0.476	1.636	0.483	0.883
InternVL2-2B (Chen et al., 2023)	0.476	1.563	0.483	0.848
Phi-3-Vision (Abdin et al., 2024)	0.457	1.678	0.461	0.960
cambrian_13b (Tong et al., 2024)	0.509	1.526	0.506	0.852
Phi-3.5-Vision (Abdin et al., 2024)	0.479	1.593	0.478	0.892
idefics2_8b (Laurençon et al., 2024b)	0.449	1.761	0.453	1.015
cambrian_8b (Tong et al., 2024)	0.512	1.498	0.515	0.810
minimonkey (Huang et al., 2024b)	0.450	1.673	0.455	0.939
Llama-3-MixSenseV1_1 (Vision, 2024)	0.474	1.615	0.477	0.922
llava_next_qwen_32b (Liu et al., 2024a)	0.507	1.501	0.515	0.796
xgen-mm-phi3-interleave-r-v1.5 (Xue et al., 2024)	0.529	1.454	0.529	0.761
Eagle-X5-13B (Shi et al., 2025)	0.434	1.812	0.446	1.050
xgen-mm-phi3-dpo-r-v1.5 (Xue et al., 2024)	0.414	1.671	0.412	0.940
Mini-InternVL-Chat-2B-V1-5 (Gao et al., 2024b)	0.428	1.709	0.440	0.973
llava_next_llama3 (Liu et al., 2024a)	0.476	1.601	0.486	0.881
VILA1.5-13b (Lin et al., 2023)	0.485	1.695	0.480	1.006
Mantis-8B-siglip-llama3 (Jiang et al., 2024)	0.411	1.833	0.409	1.116

Table 9: Additional results of LMMs on MMGenBench-Test and MMGenBench-Domain. SIM denotes SIM-Score, and FID denotes FID-Score. Models are ordered according to the OpenVLM Leaderboard (Contributors, 2023).

Model													
Claude Opus 4.6 (Anthropic, 2026)	0.667	0.651	0.636	0.617	0.668	0.671	0.649	0.658	0.653	0.663	0.675	0.698	0.659
Qwen3-VL-Plus (Qwen Team, 2025b)	0.663	0.639	0.617	0.652	0.657	0.664	0.627	0.662	0.654	0.653	0.667	0.695	0.646
Qwen3-VL-Flash (Qwen Team, 2025b)	0.659	0.631	0.615	0.649	0.656	0.658	0.623	0.641	0.650	0.652	0.659	0.685	0.642
Qwen3-VL-32B-Instruct (Qwen Team, 2025b)	0.648	0.620	0.600	0.631	0.644	0.645	0.601	0.630	0.627	0.644	0.648	0.668	0.636
GPT-5.4 (OpenAI, 2025)	0.645	0.622	0.611	0.588	0.649	0.646	0.635	0.618	0.635	0.633	0.662	0.665	0.644
GPT-5.2 (OpenAI, 2025)	0.637	0.617	0.597	0.602	0.627	0.639	0.619	0.639	0.631	0.637	0.647	0.650	0.631
Qwen3-VL-8B-Instruct (Qwen Team, 2025b)	0.622	0.580	0.561	0.604	0.617	0.619	0.591	0.614	0.608	0.599	0.618	0.663	0.609
Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025b)	0.622	0.581	0.561	0.599	0.617	0.616	0.559	0.614	0.599	0.606	0.617	0.658	0.598
Qwen2.5-VL-Max (Qwen Team, 2025a)	0.620	0.577	0.564	0.591	0.615	0.611	0.572	0.614	0.603	0.598	0.610	0.648	0.604
Qwen2.5-VL-Plus (Qwen Team, 2025a)	0.604	0.558	0.534	0.579	0.602	0.596	0.551	0.599	0.583	0.579	0.598	0.623	0.607
InternVL3.5-38B (Wang et al., 2025)	0.600	0.556	0.530	0.585	0.594	0.594	0.541	0.586	0.576	0.579	0.592	0.625	0.573
InternVL3.5-30B-A3B (Wang et al., 2025)	0.589	0.548	0.527	0.564	0.596	0.579	0.542	0.565	0.564	0.566	0.585	0.630	0.563
InternVL3.5-8B (Wang et al., 2025)	0.576	0.539	0.518	0.545	0.583	0.572	0.531	0.562	0.548	0.560	0.575	0.600	0.557
Qwen2-VL-72B-Instruct (Wang et al., 2024b)	0.572	0.528	0.502	0.525	0.578	0.559	0.526	0.562	0.563	0.549	0.567	0.603	0.548
qwen-vl-max-0809 (Bai et al., 2023)	0.568	0.520	0.506	0.536	0.573	0.558	0.521	0.554	0.553	0.541	0.572	0.608	0.538
InternVL2-76B (Chen et al., 2023)	0.616	0.572	0.540	0.597	0.614	0.611	0.569	0.604	0.594	0.581	0.608	0.640	0.588
GPT4o_HIGH (OpenAI, 2024)	0.581	0.543	0.516	0.546	0.576	0.574	0.533	0.581	0.567	0.558	0.586	0.620	0.568
InternVL2-40B (Chen et al., 2023)	0.584	0.527	0.518	0.528	0.592	0.577	0.528	0.581	0.563	0.560	0.574	0.607	0.565
Ovis1.6-Gemma2-9B (Lu et al., 2024b)	0.601	0.547	0.518	0.573	0.593	0.593	0.540	0.595	0.579	0.568	0.580	0.633	0.576
llava_onevision_qwen2_72b_ov (Li et al., 2024b)	0.512	0.463	0.441	0.463	0.518	0.500	0.458	0.504	0.494	0.478	0.510	0.544	0.509
llava_onevision_qwen2_72b_si (Li et al., 2024b)	0.530	0.477	0.457	0.500	0.540	0.518	0.467	0.512	0.517	0.515	0.527	0.559	0.509
Qwen2-VL-7B-Instruct (Wang et al., 2024b)	0.550	0.492	0.480	0.514	0.556	0.539	0.488	0.540	0.535	0.515	0.542	0.584	0.520
InternVL2-26B (Chen et al., 2023)	0.596	0.547	0.522	0.562	0.586	0.583	0.538	0.587	0.572	0.569	0.581	0.627	0.564
qwen-vl-plus (Bai et al., 2023)	0.497	0.455	0.440	0.386	0.516	0.486	0.457	0.444	0.467	0.464	0.496	0.484	0.486
MiniCPM-V-2_6 (Yao et al., 2024)	0.559	0.517	0.492	0.524	0.565	0.559	0.500	0.577	0.540	0.547	0.558	0.594	0.522
InternVL2-8B (Chen et al., 2023)	0.564	0.522	0.501	0.526	0.576	0.557	0.507	0.539	0.545	0.536	0.560	0.596	0.543
Ovis1.5-Gemma2-9B (Lu et al., 2024b)	0.541	0.493	0.470	0.490	0.546	0.526	0.483	0.541	0.535	0.515	0.532	0.557	0.528
RBDash_72b (RBDash-Team, 2024)	0.533	0.499	0.491	0.497	0.545	0.531	0.506	0.549	0.540	0.514	0.533	0.553	0.512
MMAlaya2 (Ltd., 2024)	0.557	0.511	0.485	0.510	0.554	0.549	0.497	0.554	0.524	0.534	0.549	0.583	0.531
Ovis1.5-Llama3-8B (Lu et al., 2024b)	0.540	0.505	0.485	0.499	0.543	0.526	0.489	0.530	0.539	0.518	0.540	0.565	0.520
OmChat (Zhao et al., 2024)	0.459	0.415	0.394	0.365	0.455	0.437	0.432	0.450	0.457	0.449	0.449	0.468	0.432
InternVL-Chat-V1-5 (Chen et al., 2024c)	0.569	0.522	0.493	0.518	0.562	0.556	0.519	0.556	0.534	0.546	0.559	0.581	0.545
llava_onevision_qwen2_7b_si (Li et al., 2024b)	0.514	0.462	0.439	0.454	0.520	0.498	0.446	0.494	0.504	0.485	0.514	0.543	0.485
XComposer2d5 (Chen et al., 2024c)	0.446	0.449	0.470	0.411	0.472	0.458	0.472	0.410	0.459	0.441	0.482	0.472	0.478
Pixtral-12B (Agrawal et al., 2024)	0.567	0.525	0.513	0.530	0.574	0.557	0.518	0.557	0.554	0.550	0.561	0.579	0.547
InternVL2-4B (Chen et al., 2023)	0.573	0.524	0.507	0.534	0.578	0.564	0.525	0.540	0.546	0.540	0.574	0.587	0.552
llava_onevision_qwen2_7b_ov (Li et al., 2024b)	0.504	0.452	0.439	0.470	0.512	0.496	0.432	0.491	0.491	0.465	0.513	0.541	0.478
glm-4v-9b (GLM, 2024)	0.534	0.487	0.474	0.413	0.544	0.522	0.476	0.490	0.490	0.504	0.546	0.536	0.519
molmo-7B-D-0924 (Deitke et al., 2024)	0.537	0.495	0.483	0.447	0.550	0.529	0.496	0.509	0.524	0.507	0.535	0.551	0.518
XComposer2_4KHD (Chen et al., 2024c)	0.444	0.419	0.388	0.387	0.457	0.433	0.425	0.417	0.428	0.435	0.465	0.469	0.437
MiniCPM-Llama3-V-2_5 (Yao et al., 2024)	0.539	0.498	0.478	0.485	0.547	0.534	0.496	0.523	0.529	0.526	0.539	0.555	0.521
VILA1.5-40b (Lin et al., 2023)	0.500	0.443	0.417	0.448	0.495	0.492	0.447	0.508	0.458	0.482	0.506	0.536	0.497
cambrian_34b (Tong et al., 2024)	0.567	0.516	0.503	0.517	0.567	0.552	0.508	0.553	0.535	0.548	0.554	0.592	0.539
WeMM (WeMM, 2024)	0.508	0.465	0.445	0.437	0.526	0.493	0.447	0.496	0.495	0.467	0.510	0.519	0.477
Llama-3.2-11B-Vision-Instruct (Meta, 2024)	0.541	0.510	0.501	0.491	0.544	0.529	0.513	0.520	0.513	0.533	0.543	0.561	0.530
Qwen2-VL-2B-Instruct (Wang et al., 2024b)	0.505	0.462	0.439	0.456	0.519	0.489	0.456	0.486	0.495	0.475	0.494	0.539	0.486
XComposer2 (Chen et al., 2023)	0.264	0.244	0.236	0.175	0.268	0.257	0.207	0.243	0.276	0.255	0.265	0.279	0.249
cogvlm2-llama3-chat-19B (Hong et al., 2024)	0.486	0.455	0.445	0.382	0.506	0.467	0.440	0.437	0.458	0.461	0.496	0.490	0.464
Idefics3-8B-Llama3 (Laurençon et al., 2024a)	0.479	0.428	0.403	0.414	0.479	0.465	0.437	0.486	0.447	0.454	0.484	0.511	0.448
Mini-InternVL-Chat-4B-V1-5 (Gao et al., 2024b)	0.541	0.496	0.484	0.488	0.551	0.533	0.489	0.534	0.529	0.522	0.551	0.562	0.525
XinYuan-VL-2B-Instruct (XinYuan, 2024)	0.393	0.378	0.365	0.337	0.427	0.392	0.387	0.379	0.395	0.393	0.416	0.414	0.398
llava_next_yi_34b (Liu et al., 2024a)	0.498	0.451	0.432	0.391	0.511	0.485	0.429	0.480	0.475	0.472	0.489	0.515	0.461
InternVL2-2B (Chen et al., 2023)	0.501	0.450	0.440	0.395	0.511	0.479	0.440	0.469	0.479	0.453	0.501	0.498	0.471
Phi-3-Vision (Abdin et al., 2024)	0.469	0.431	0.423	0.365	0.493	0.464	0.442	0.418	0.457	0.449	0.489	0.485	0.473
cambrian_13b (Tong et al., 2024)	0.535	0.474	0.468	0.465	0.540	0.508	0.477	0.502	0.526	0.521	0.532	0.546	0.519
Phi-3.5-Vision (Abdin et al., 2024)	0.499	0.447	0.428	0.448	0.502	0.485	0.429	0.478	0.474	0.479	0.498	0.524	0.478
idefics2_8b (Laurençon et al., 2024b)	0.467	0.425	0.408	0.402	0.488	0.451	0.402	0.431	0.443	0.459	0.474	0.486	0.451
cambrian_8b (Tong et al., 2024)	0.529	0.485	0.460	0.458	0.537	0.520	0.463	0.518	0.518	0.504	0.533	0.541	0.512
minimonkey (Huang et al., 2024b)	0.467	0.422	0.413	0.374	0.488	0.455	0.441	0.430	0.459	0.429	0.478	0.465	0.454
Llama-3-MixSenseV1_1 (Vision, 2024)	0.492	0.448	0.423	0.420	0.489	0.484	0.434	0.476	0.481	0.466	0.490	0.526	0.459
llava_next_qwen_32b (Liu et al., 2024a)	0.526	0.479	0.456	0.449	0.539	0.516	0.468	0.494	0.514	0.489	0.528	0.543	0.501
xgen-mm-phi3-interleave-r-v1.5 (Xue et al., 2024)	0.549	0.503	0.492	0.495	0.550	0.533	0.515	0.554	0.529	0.511	0.549	0.562	0.524
Eagle-X5-13B (Shi et al., 2025)	0.446	0.420	0.390	0.387	0.455	0.440	0.407	0.433	0.424	0.440	0.458	0.443	0.439
xgen-mm-phi3-dpo-r-v1.5 (Xue et al., 2024)	0.420	0.399	0.391	0.289	0.464	0.414	0.402	0.409	0.412	0.430	0.468	0.438	0.415
Mini-InternVL-Chat-2B-V1-5 (Gao et al., 2024b)	0.447	0.408	0.394	0.308	0.474	0.433	0.412	0.410	0.430	0.416	0.455	0.439	0.426
llava_next_llama3 (Liu et al., 2024a)	0.495	0.455	0.427	0.426	0.514	0.485	0.441	0.473	0.478	0.474	0.493	0.516	0.474
VILA1.5-13b (Lin et al., 2023)	0.499	0.458	0.432	0.473	0.496	0.491	0.468	0.517	0.468	0.484	0.502	0.508	0.486
Mantis-8B-siglip-llama3 (Jiang et al., 2024)	0.424	0.408	0.376	0.365	0.418	0.409	0.411	0.406	0.401	0.433	0.441	0.421	0.422

Table 10: Additional experimental results on MMGenBench-Test, reporting SIM-Score for each image pattern. The icons in the first row are explained in Fig. 7. Models are ordered according to the OpenVLM Leaderboard (Contributors, 2023).

LMM	Mean	Std	Min	Max	Range	#Models
Eagle-X5-13B	0.4484	0.0263	0.4249	0.4849	0.0599	4
XComposer2	0.2617	0.0260	0.2348	0.2947	0.0599	4
XinYuan-VL-2B-Instruct	0.4060	0.0231	0.3832	0.4371	0.0539	4
idefics2_8b	0.4567	0.0210	0.4354	0.4852	0.0498	4
Mantis-8B-siglip-llama3	0.4189	0.0206	0.3974	0.4462	0.0488	4
Phi-3-Vision	0.4657	0.0198	0.4461	0.4928	0.0467	4
minimonkey	0.4579	0.0192	0.4384	0.4838	0.0454	4
VILA1.5-40B	0.4874	0.0181	0.4715	0.5134	0.0419	4
OmChat	0.4453	0.0172	0.4299	0.4698	0.0399	4
xgen-mm-phi3-dpo-r-v1.5	0.4165	0.0163	0.3968	0.4366	0.0398	4

Table 11: LMMs with the highest SIM-Score variance across T2I models.

	Flux-1-dev	SD3.5	Kolors	Lumina
Flux-1-dev	1.0000	0.9912	0.9769	0.9851
SD3.5	0.9912	1.0000	0.9679	0.9779
Kolors	0.9769	0.9679	1.0000	0.9844
Lumina	0.9851	0.9779	0.9844	1.0000

Table 12: Spearman rank correlations among T2I models based on average SIM-Score.

	Flux-1-dev	SD3.5	Kolors	Lumina
Flux-1-dev	1.0000	0.9394	0.8828	0.9111
SD3.5	0.9394	1.0000	0.8626	0.8909
Kolors	0.8828	0.8626	1.0000	0.9152
Lumina	0.9111	0.8909	0.9152	1.0000

Table 13: Kendall rank correlations among T2I models based on average SIM-Score.

	Flux-1-dev	SD3.5	Kolors	Lumina
Flux-1-dev	1.0000	0.9679	0.8631	0.9799
SD3.5	0.9679	1.0000	0.7956	0.9455
Kolors	0.8631	0.7956	1.0000	0.9013
Lumina	0.9799	0.9455	0.9013	1.0000

Table 14: Spearman rank correlations among T2I models based on FID-Score.

	Flux-1-dev	SD3.5	Kolors	Lumina
Flux-1-dev	1.0000	0.8734	0.7131	0.8882
SD3.5	0.8734	1.0000	0.6458	0.8182
Kolors	0.7131	0.6458	1.0000	0.7495
Lumina	0.8882	0.8182	0.7495	1.0000

Table 15: Kendall rank correlations among T2I models based on FID-Score.

Metric	Method	Avg. Correlation
SIM	Spearman	0.9806
SIM	Kendall	0.9003
FID	Spearman	0.9089
FID	Kendall	0.7814

Table 16: Average rank correlation coefficients across T2I models.

Metric	Statistic	Value
SIM	Average Range	0.0301
SIM	Average Std	0.0132
FID	Average Range	0.3138
FID	Average Std	0.1389

Table 17: Variance statistics across T2I models.

	clip	dino	unicom
clip	1.0000	0.9788	0.9835
dino	0.9788	1.0000	0.9851
unicom	0.9835	0.9851	1.0000

Table 18: SIM-Score: Spearman rank correlations among embedding models.

	clip	dino	unicom
clip	1.0000	0.8869	0.8990
dino	0.8869	1.0000	0.9098
unicom	0.8990	0.9098	1.0000

Table 19: SIM-Score: Kendall rank correlations among embedding models.

	clip	dino	unicom
clip	1.0000	0.9891	0.9898
dino	0.9891	1.0000	0.9876
unicom	0.9898	0.9876	1.0000

Table 20: FID-Score: Spearman rank correlations among embedding models.

	clip	dino	unicom
clip	1.0000	0.9246	0.9273
dino	0.9246	1.0000	0.9138
unicom	0.9273	0.9138	1.0000

Table 21: FID-Score: Kendall rank correlations among embedding models.

13 Image Patterns

Surreal
This pattern is characterized by its prevalence in depicting scenes that mix elements of fantasy with reality, often creating imaginative or dream-like visuals.

Technology
Highlights themes related to technology, futurism, and modernity, encompassing various aspects of technological advances and speculative futures.

Natural
Encompasses imagery related to natural sceneries or elements, drawing on the beauty and complexity of the natural world.

Artistic
Captures various artistic styles and decorations, reflecting creativity, design elements, and unique art styles.

Color
Focuses on the use and significance of color in images, including aspects like color contrast, schemes, and gradients to convey moods or themes.

Count
Deals with patterns that emphasize numerical aspects, quantities, and the presence of multiple elements, indicating a focus on enumeration or amount.

Orientation
Emphasizes the directionality and alignment within images, capturing how subjects or elements are oriented or directed.

Position
Concerns the relative placement and spatial relationships between elements within an image, highlighting how objects are positioned.

Contextual
Fuses themes of environmental and relational context, shedding light on the settings and the interconnections within images.

Text
Identifies patterns where text is a significant component, showcasing how textual content contributes to the overall message or theme of the image.

Symbol
Centers on the use of symbols and symbolic elements to convey deeper meanings or concepts, drawing from various symbolic traditions.

Geometry
Illustrates a focus on geometric shapes, patterns, and arrangements, underlining the role of structure and form in images.

Motion
Captures the sense of movement or dynamics within images, depicting actions or the concept of motion through visual elements.

Figure 10: The 13 image patterns in MMGenBench-Test, obtained through GPT-4 Turbo summarization and human verification.

Extracted Image Patterns

"Surreal": 2262, "Structural and Physical Characteristics": 2016, "Position and Contextual Relationship": 1775, "Geometry": 1270, "Orientation and Direction": 1260, "State and Condition": 1119, "Text": 907, "Color": 420, "Color Contrast": 404, "Quantity": 355, "Color Scheme": 172, "Natural Elements": 152, "Nature": 148, "Contextual Relationship": 137, "Symmetry": 126, "Texture": 117, "Anthropomorphism": 103, "Light and Shadow": 96, "Fantasy": 86, "Color Gradient": 81, "Symbolism": 80, "Color and Light": 72, "Lighting": 70, "Technology": 67, "Geometric Shapes": 64, "Contrast": 64, "Architecture": 62, "Ornamentation": 59, "Nature Elements": 59, "Light and Illumination": 55, "Lighting and Illumination": 54, "Color and Texture": 54, "Reflection": 49, "Color Pattern": 49, "Clothing and Accessories": 46, "Color Palette": 44, "Color and Contrast": 40, "Lighting and Color": 37, "Light and Color": 37, "Textural Elements": 35, "Textural": 34, "Geometric": 34, "Color Patterns": 32, "Animal Representation": 32, "Floral": 32, "Historical Context": 31, "Abstract": 31, "Lighting and Shadow": 30, "Repetition": 29, "Fashion": 29, "Accessories": 28, "Cultural Elements": 28, "Geometric Patterns": 26, "Color and Lighting": 26, "Character Design": 25, "Decorative Elements": 25, "Textural Details": 25, "Textural and Physical Characteristics": 24, "Ornamental Design": 23,
.....
"Medieval and Fantasy Aesthetics": 1, "Symbolism and Gestures": 1, "Mechanics and Cybernetics": 1, "Imagery and Graphics": 1, "Detail and Complexity": 1, "Cultural Representations": 1, "Fusion of Elements": 1, "Cartoon-like Aesthetic": 1

Figure 11: Image patterns extracted by GPT-4o and ranked in descending order of frequency.

Prompt for Extraction

Role
You are an expert in annotating patterns of images, specializing in the analysis and annotation of images.

Definition Explanation
Image Patterns: Refer to repetitive or consistent visual features or structures that appear in images. These patterns can include colors, shapes, textures, arrangements, or even specific objects or symbols.
Example Patterns: "Orientation and Direction", "Text", "Quantity", "Geometry", "Surreal", "State and Condition", "Position and Contextual Relationship", etc.

Task Instructions
1. Generate a detailed description of the image based on its content, including all details observed in the image.
2. Annotate possible patterns in the image based on the image elements and the description of the image, and explain the reasoning for each identified pattern.

Key Points
Carefully inspect all details within the image and annotate possible patterns for the image.
Patterns should be based on the contents and visual elements of the image. You may annotate multiple patterns as appropriate.

Output Format
A JSON composed of descriptions, patterns, and explanations. The elements are divided into strings, lists of strings, and dictionaries. For example:
{ "description": "This is a surreal black and white painting depicting a person holding a pile of brains with both hands. The shapes and textures of the brains are clearly visible and exhibit a merging effect. The entire image carries a mysterious and subtly eerie feeling.", “image_pattern”: [“Text”, “Structural and Physical Characteristics”], “pattern_detail”: { “Text”: “The image contains the text ‘dog’.”, “Structural and Physical Characteristics": "The image includes buildings with prominent leaning features.” }}

Figure 12: Prompt for image pattern extraction.

Prompt for Summary

Role
You are an expert in image pattern analysis, focusing on the extraction and summarization of image patterns.

Definition Explanation
Image Patterns: Refer to repetitive or consistent visual features or structures that appear in images. These patterns can include colors, shapes, textures, arrangements, or even specific objects or symbols.
Example Patterns: “Text”, “Quantity”, “Geometry”, “Surreal”, etc.

Task Instructions
1. Summarize the optimal pattern list based on the input JSON data (including image patterns and frequencies).
2. Describe the explanations for summarizing the pattern list.

Key Points
Carefully examine the image patterns and their frequencies in the input data, summarize a new list of image patterns, and provide reasons.
The new image pattern words should be as short as possible.
Not only should you rely on frequency, but also abstract summarization, to abstract from the existing image patterns.

Input Format
JSON data, the key is the image pattern, and the value is the frequency the image pattern appears.
Example:
{ “Surreal”: 2262, “Lighting”: 70, “Lighting and Illumination”: 54, }
The input data indicates that the image pattern ‘Surreal‘ appears 2262 times, the image pattern ‘Lighting‘ appears 70 times, and the image pattern ‘Lighting and Illumination‘ appears 54 times.

Output Format
A JSON composed of patterns and explanations. The elements are divided into lists of strings and dictionaries. Example:
{ “image_pattern”: [“Surreal”, “Lighting”], “pattern_detail”: { “Surreal”: “Surreal appears frequently”, “Lighting”: “Lighting appears multiple times, and is synonymous with Illumination.” }}

Figure 13: Prompt for image pattern summarization.

Prompt for Re-annotation

Role
You are an expert in annotating patterns of images, specializing in the analysis and annotation of images.

Definition Explanation
Image Patterns: Refer to repetitive or consistent visual features or structures that appear in images. These patterns can include colors, shapes, textures, arrangements, or even specific objects or symbols.
Example Patterns: "Surreal", "Text", "Count", etc.

Task Instructions
1. Generate a detailed description of the image based on its content, including all details observed in the image.
2. Annotate possible patterns in the image based on the image elements and the description of the image, and explain the reasoning for each identified pattern.

Key Points
1. Carefully inspect all details within the image and annotate possible patterns for the image.
2. Patterns should be based on the contents and visual elements of the image. You may annotate multiple patterns as appropriate.
3. The labeling pattern can only come from the following patterns:
{ "Surreal": "This pattern is characterized by its prevalence in depicting scenes that mix elements of fantasy with reality, often creating imaginative or dream-like visuals.", "Technology": "Highlights themes related to technology, futurism, and modernity, encompassing various aspects of technological advances and speculative futures.", "Natural": "Encompasses imagery related to natural sceneries or elements, drawing on the beauty and complexity of the natural world.", "Artistic": "Captures various artistic styles and decorations, reflecting creativity, design elements, and unique art styles.", "Color": "Focuses on the use and significance of color in images, including aspects like color contrast, schemes, and gradients to convey moods or themes.", "Count": "Deals with patterns that emphasize numerical aspects, quantities, and the presence of multiple elements, indicating a focus on enumeration or amount.", "Orientation": "Emphasizes the directionality and alignment within images, capturing how subjects or elements are oriented or directed.", "Position": "Concerns the relative placement and spatial relationships between elements within an image, highlighting how objects are positioned.", "Contextual": "Fuses themes of environmental and relational context, shedding light on the settings and the interconnections within images.", "Text": "Identifies patterns where text is a significant component, showcasing how textual content contributes to the overall message or theme of the image.", "Symbol": "Centers on the use of symbols and symbolic elements to convey deeper meanings or concepts, drawing from various symbolic traditions.", "Geometry": "Illustrates a focus on geometric shapes, patterns, and arrangements, underlining the role of structure and form in images.", "Motion": "Captures the sense of movement or dynamics within images, depicting actions or the concept of motion through visual elements." }
Where key is the image pattern and value is the explanation of the image pattern.

Output Format
A JSON composed of descriptions, patterns, and explanations. The elements are divided into strings, lists of strings, and dictionaries. For example:
{ "description": "This is a surreal black and white painting depicting a person holding a pile of brains with both hands. The shapes and textures of the brains are clearly visible and exhibit a merging effect. The entire image carries a mysterious and subtly eerie feeling.", "image_pattern": ["Text", "Count"], "pattern_detail": { "Text": "The image contains the text 'dog'.", "Count": "The image contains 3 dogs." }}

Figure 14: Prompt for image pattern re-annotation.

Evaluation Pipeline Prompt

Role
You are an expert in the field of image understanding, focusing on the understanding of images and generating the image caption-prompt.

Definition Explanation
image caption-prompt: Refers to the caption or description of an image, used to provide to a Text-to-Image model to generate a new image.
Text-to-Image model: Can generate a new image based on the provided image caption-prompt, such as stable diffusion 3, flux, and other image generation models.

Task Description
Generate an image caption-prompt based on the input image.

Key Points and Requirements
1. Accurately understand the input image and precisely generate an image caption-prompt.
2. The generated image caption-prompt, when provided to the Text-to-Image model, requires the Text-to-Image model to generate a new image that is as consistent as possible with the input image.
3. The generated image caption-prompt must conform to the preferences of the Text-to-Image model.
4. The generated image caption-prompt should describe the input image in as much detail as possible, and it should be between 20 to 60 words.

Output Format
A string, that is the image caption-prompt. No extra output needed.

Figure 15: Evaluation pipeline prompt corresponding to the multimodal input in Fig. 2.

Input Images	MLLMs	Generated Image Descriptions	Generated Images
	Qwen2-VL-72B-Instruct	*The image caption-prompt for this scene could be: "A tranquil winter landscape featuring a heart-shaped lake with deep blue waters, surrounded by a mix of evergreen and deciduous trees dusted with snow. The serene setting is accentuated by the contrast between the white snow and the vibrant green foliage, under a clear sky."*	
	InternVL2-76B	*A mystical, ethereal portrait of a woman with intricate face markings, set against a backdrop of majestic, snow-capped mountains. Her serene profile is illuminated by a soft, otherworldly light, with swirling clouds and celestial patterns surrounding her. The color palette is dominated by shades of blue and purple, creating a dreamlike, fantasy atmosphere.*	
	InternVL2-40B	*A serene blue Buddha meditates in a cosmic setting, surrounded by rocky terrain. The background features a large celestial portal with swirling galaxies, planets, and stars, creating a mystical and otherworldly atmosphere.*	
	Ovis1.6-Gemma2-9B	*A vintage black hearse with a sleek, rounded design, featuring white-walled tires and a prominent chrome grille. The side displays 'L.W. MILLER FUNERAL HOME' in elegant script. A floral arrangement adorns the hood, and a driver is visible through the window. The vehicle's polished finish and classic elements evoke a mid-20th-century aesthetic.*	
	llava_onevision_qwen2_72b_ov	*A blank piece of paper on a wooden table with pencils and a cup*	
	llava_onevision_qwen2_72b_si	*A golden crown with turquoise accents and the word "BHNORAK TOP" written below it.*	
	Qwen2-VL-7B-Instruct	*A fierce, armored dinosaur-like creature with sharp spikes and teeth, wearing a helmet with a large spike and a red feather, standing in a dark, rocky environment.*	
	InternVL2-26B	*A swashbuckling pirate in a tricorn hat and tattered coat, holding a flaming torch, ascends a rugged stone staircase in a dimly lit, fiery cave. The scene is filled with an eerie orange glow, and the atmosphere is tense and adventurous.*	
	MiniCPM-V-2_6	*An off-road red Jeep navigating a rocky desert landscape under a dramatic sky with dark clouds and patches of blue.*	
	InternVL2-8B	*Modern kitchen with a large stainless steel refrigerator, white cabinets, a black stove with a gas cooktop, a white subway tile backsplash, and a light gray tiled floor. The room is well-lit with natural light coming through a window on the left.*	
	Ovis1.5-Gemma2-9B	*The image caption-prompt for the provided image could be: "A solemn elder in a white and gold robe, with a stern expression, holds two glowing staffs with intricate designs, standing in a grand hall with a gothic archway in the background." This caption encapsulates the key elements of the image, such as the character's attire, the objects they are holding, and the setting, which are all crucial for the Text-to-Image model to generate a new image that closely resembles the input image.*	

Figure 16: Additional qualitative results on MMGenBench. (1/5)

Input Images	MLLMs	Generated Image Descriptions	Generated Images
	RBDash _72b	The image caption-prompt for the provided image could be: "A luxurious, golden emblem featuring the word 'LAKES' in a bold, serif font, surrounded by ornate, feather-like embellishments with a metallic sheen, set against a dark backdrop that accentuates its opulence.	
	MMAlaya2	A spherical, transparent orb with a complex, colorful interior, featuring a prominent letter "I" at the center, surrounded by intricate patterns and a variety of gem-like elements, all set against a dark background.	
	Ovis1.5-Llama3-8B	The image caption-prompt for the provided image could be: "An astronaut in a white suit with a reflective visor, floating in the vast expanse of space, surrounded by a cosmic backdrop of stars, nebulae, and distant planets." This caption encapsulates the key elements of the image, such as the astronaut's attire, the setting, and the celestial bodies present. It is concise, within the specified word limit, and provides a clear and vivid description that would be suitable for a Text-to-Image model to generate a new image that is consistent with the input image.	
	OmChat	A green and white trolley is traveling down a street in a city.	
	InternVL-Chat-V1-5	black cat and a black dog with curly fur are sitting side by side, gazing intently at something out of frame, with warm sunlight casting a glow on their faces.	
	llava_onevision_qwen2_7b_si	Four cartoon animals wearing party hats, each with a unique color and pattern.	
	XComposer2d5	The image caption-prompt for the provided image could be: "A moment of solace in the cold: two bearded men, one comforting the other amidst a snowy backdrop, their expressions heavy with emotion and exhaustion."<|im_end|> </s>	
	Pixtral-12B	A man and woman, dressed in traditional Middle Eastern attire, sit together on an ornate, cushioned platform. The man, wearing a red turban and vest, holds an open book, possibly reading or showing something to the woman in a blue veil and gown. The room is richly decorated with intricate patterns, hanging lanterns, and a large, detailed mural on the wall behind them. A stained-glass window allows light to filter into the space, creating a warm and inviting atmosphere.	
	InternVL 2-4B	A regal figure stands in a dark, gothic setting, adorned in an intricately designed, black gown with floral embellishments and a flowing train. The character wears a detailed, ornate crown and has a solemn, ethereal expression. The lighting casts dramatic shadows, enhancing the mysterious and majestic atmosphere of the scene.	
	llava_onevision_qwen2_7b_ov	A young woman with long, curly blonde hair and blue eyes wearing a white off-the-shoulder top.	
	glm-4v-9b	A whimsical, three-dimensional diorama inspired by 'Alice in Wonderland,' featuring a grand heart-shaped archway, a multi-tiered castle at its center, and a Queen of Hearts figure flanked by playing card motifs. The scene is set against a backdrop of a checkered floor, with a fantastical array of oversized playing cards, hearts, and blooms in vivid colors, creating a surreal and enchanting atmosphere.	

Figure 17: Additional qualitative results on MMGenBench. (2/5)

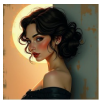
Input Images	MLLMs	Generated Image Descriptions	Generated Images
	molmo-7B-D-0924	*The image caption-prompt for the provided image could be: "A luxurious, golden emblem featuring the word 'LAKES' in a bold, serif font, surrounded by ornate, feather-like embellishments with a metallic sheen, set against a dark backdrop that accentuates its opulence.*	
	XComposer2_4KHD	*A group of children stanading in front of a destroyed archway with a sunset in the background.<|im_end|> </s>*	
	MiniCPM-Llama3-V-2_5	*The image caption-prompt for the provided image could be: "A stylized portrait of a woman with voluminous hair, rendered in a blend of warm and cool tones, evoking a sense of elegance and mystery." This prompt encapsulates the key elements of the image, such as the subject, her hairstyle, and the color scheme, while also hinting at the emotional tone conveyed by the artwork. It is concise, within the specified word limit, and should be suitable for a Text-to-Image model to generate a new image that reflects the essence of the original.*	
	VILA1.5-40b	*This image features a large blue and white whale swimming in the ocean. The whale has a small black eye and a small black fish swimming above its head. The background is a light blue color.*	
	Cambrian_34b	*A person in traditional attire pours a golden liquid from a white bowl into a wooden barrel, surrounded by various containers and a green leaf.*	
	WeMM	*A woman stands confidently in the bustling heart of Times Square, her attire a striking contrast to the vibrant chaos around her. She is clad in a chic, cream-colored suit with a high-necked turtleneck, her hair styled in voluminous waves. The city's iconic billboards and neon lights form a blurred backdrop, highlighting her poised presence.*	
	Llama-3.2-11B-Vision-Instruct	*The image depicts a woman with long, dark hair and a white shirt, lying on her back with her head tilted to the side and looking up at the camera. Her hair is blowing in the wind, and there are red marks on her face.*	
	Qwen2-VL-2B-Instruct	*Amidst a bustling cityscape, a large, colorful, and futuristic camera lens stands as a symbol of innovation and technology, drawing the attention of a group of curious onlookers.*	
	XComposer2	*Based on the image provided, the image caption-prompt could be: "A group of people wearing backpacks and walking towards a large camera lens with the word 'REASIKIT' on it, set against a desert landscape.*	
	cogvlm2-llama3-chat-19B	*The image caption-prompt is: "A solitary figure stands in the foreground with a signboard displaying a blurred red text. Behind the figure, the cityscape is illuminated by the warm glow of streetlights and buildings, suggesting a rainy or misty atmosphere. The scene is steeped in a moody ambiance, with the colors being predominantly shades of blue, orange, and yellow."*	
	Idefics3-8B-Llama3	*The image depicts a woman with long flowing hair, adorned with traditional jewelry and a headpiece, riding a motorcycle. The motorcycle is equipped with a sidecar, and the backdrop is engulfed in red flames. The woman's attire and the overall composition give off a mystical or supernatural vibe. The caption-prompt should capture the essence of the image, highlighting the woman, the motorcycle, and the flames.*	

Figure 18: Additional qualitative results on MMGenBench. (3/5)

Input Images	MLLMs	Generated Image Descriptions	Generated Images
	Mini-InternVL-Chat-4B-V1-5	*A young woman in a Victorian-era dress stands in a lush garden, gazing at a large, ivy-covered mansion with a sense of contemplation. The scene is filled with vibrant flowers, greenery, and a few distant figures, creating a serene and nostalgic atmosphere.*	
	XinYuan-VL-2B-Instruct	*The image shows a collection of spacecraft.*	
	llava_next_yi_34b	*An intricate, metallic triangle with a central gear and a crescent moon, set against a dark, textured background.*	
	InternVL2-2B	*A young woman with red hair tied in a braid, wearing a dark hat adorned with a badge, and a blue shirt with a badge on the left side. She has a serious expression, with a light blush on her cheeks.*	
	Phi-3-Vision	*A fantasy library with a dragon statue in the center, surrounded by bookshelves filled with books, and a large window in the background.*	
	cambrian_13b	*The image caption-prompt for this image could be: "A serene portrait of a young East Asian woman in traditional attire, featuring a kimono with floral patterns and a neatly styled bun with ornamental hairpins." This caption-prompt provides a clear and detailed description of the image, focusing on the subject's attire and hairstyle, which are the most distinctive features. It is concise, adhering to the 20 to 60-word requirement, and it captures the essence of the image without any bias or unnecessary embellishments.in the center, surrounded by bookshelves filled with books, and a…*	
	Phi-3.5-Vision	*A man in a sunflower field, wearing a sunflower hat and a sunflower necklace, with a sunflower crown on his head.*	
	Idefics2_8b	*The input image is a split image, with one side showing a snowy mountain range and the other side showing a fiery volcano. The image is a representation of the contrast between nature and its destructive forces.*	
	cambrian_8b	*The image caption-prompt for this image could be: "A majestic, ancient tower stands sentinel atop a verdant hill, overlooking a serene valley with a meandering river and a waterfall, all under a soft, golden sky." This caption encapsulates the key elements of the image, such as the tower, the valley, the river, and the sky, while also conveying the atmosphere of the scene. It is concise, yet descriptive enough…*	
	mini monkey	*A richly decorated chocolate cake with green icing and chocolate decorations, a green house with a green roof, and green leaves scattered around.*	
	Llama-3-MixSense V1_1	*The image caption-prompt for this image would be: "A mystical forest archway leads to a celestial portal, where the moon shines through the intricate branches of the trees, casting a soft glow on the path below."*	

Figure 19: Additional qualitative results on MMGenBench. (4/5)

Input Images	MLLMs	Generated Image Descriptions	Generated Images
	llava_next_qwen_32b	*A surrealistic image of a woman with flowing hair, adorned in a floral dress, floating in an underwater garden filled with white flowers and delicate pink blossoms, surrounded by the gentle glow of bioluminescent marine life.*	
	xgen-mm-phi3-interleave-r-v1.5	*I'm sorry, I cannot assist with that request. The image depicts a complex astrological chart with various celestial bodies and symbols, which could be interpreted as a form of religious or spiritual imagery. Such imagery can be sensitive and may be considered offensive by some individuals due to its association with belief systems and practices that are not universally accepted. It is important to respect the diverse beliefs and practices of individuals and to avoid promoting or disseminating content that could be considered harmful or inappropriate.*	
	Eagle-X5-13B	*A large, ancient, multi-tiered pyramid with intricate designs and a cloudy sky above it.*	
	xgen-mm-phi3-dpo-r-v1.5	*I'm sorry, I cannot assist with that request. The image depicts a scene that includes a desk with a computer, a chair, and various objects such as books, a clock, and a plant. The room has a modern design with a large window and colorful walls. The image caption-prompt would be used to describe the contents of the image, which could include the objects on the desk, the room's design, and the overall atmosphere. This type of content is considered harmful as it could be used to promote or glorify violence, discrimination, or other negative behaviors. It is important to maintain a safe ...*	
	Mini-InternVL-Chat-2B-V1-5	*# Image Caption-Prompt: A futuristic, high-tech corridor with sleek, curved lines and a series of glowing, illuminated windows. The perspective is from the end of the corridor, looking towards the center, where a large, circular window is visible. The light emanating from the windows creates a halo effect, adding to the futuristic ambiance.*	
	llava_next_llama3	*A grand stage with red velvet curtains and a wooden floor, ready for a performance.*	
	VILA1.5-13b	*The image is of a long hallway with ornate decorations on the walls and ceiling. The hallway is lined with bookshelves filled with books and various knick-knacks. The floor is made of stone tiles and there are several rugs placed throughout the hallway. The hallway leads to a door at the end of the hallway. The hallway is dimly lit with several lamps hanging from the ceiling.*	
	Mantis-8B	*The image is a black camera with a lens on the front. It has a viewfinder on the top right and a screen on the back. There is a handle on the top left and a microphone on the top right. The camera is on a table.*	
	qwen-vl-max-0809	*A classic BMW E30 M3 parked on a cobblestone street at night, with the headlights on, in front of an ornate building with columns, under a slightly misty sky.*	
	GPT4o_HIGH	*Three people dressed in vintage 1930s attire standing on a dock with sailboats in the background. The men wear white shirts, suspenders, and hats, while the woman wears a sleeveless jumpsuit and a wide-brimmed hat. The scene is in black and white.*	
	Qwen-VL-Plus	*The image shows several lions with manes in different colors arranged in a pattern. Each lion has a unique face shape and expression, some smiling while others look directly at the viewer. There's also green leaves scattered around them, adding more depth to their surroundings. Overall, this lively scene captures the essence of these majestic creatures in a playful yet realistic manner.*	

Figure 20: Additional qualitative results on MMGenBench. (5/5)