UniReflect: Self-Reflection Tuning for Unified Multimodal Understanding and Generation

Cong Wei, Zepeng Huang, Haoxian Tan, Dashuang Liang*, Lishuai Gao*, Pengfei Yan, and Xiaoming Wei

Meituan Inc.
weicong04@meituan.com

Abstract. Recent studies indicate that test-time scaling (TTS) or reflection mechanisms can substantially improve performance in multimodal generative tasks. At the core of TTS lies verification, which assesses the semantic alignment between textual instructions and generated images. Such reflection inherently couples generation and understanding, its integration into a single unified model for visual comprehension and synthesis remains largely unexplored. Existing methods typically depend on external Visual Large Language Models (VLLMs) for verification, resulting in additional computational overhead and fragmented pipelines. We introduce UniReflect, **the first unified framework for visual understanding and generation that performs in-model verification** via self-reflection tuning. The proposed model is explicitly trained to assess the correspondence between the prompt and the generated image, produce a structured analysis, and output a special token that guides the decoding of a refined image, which enables seamless TTS within a single unified model, eliminating the need for external verifiers or extra editing prompts. Extensive experiments demonstrate that UniReflect not only advances image generation fidelity but also provides reliable visual verification, setting a new benchmark for unified multi-modal models.

Keywords: Verifier · Self-Reflection · Understanding and Generation

1 Introduction

Multi-modal generative models [1, 27, 49, 63], particularly those capable of synthesizing high-quality images from diverse textual or multimodal descriptions, have witnessed substantial advancements in recent years. Beyond improvements in visual fidelity, emerging research has increasingly focused on test-time scaling (TTS) or reflection mechanisms [20, 32, 41, 66], which aim to enhance generative outputs through integrated verification and self-assessment during the inference process. TTS strategies exploit additional reasoning or validation steps at generation time to ensure closer adherence to user specifications, thereby enabling models to iteratively refine and optimize their outputs.

* Corresponding authors.

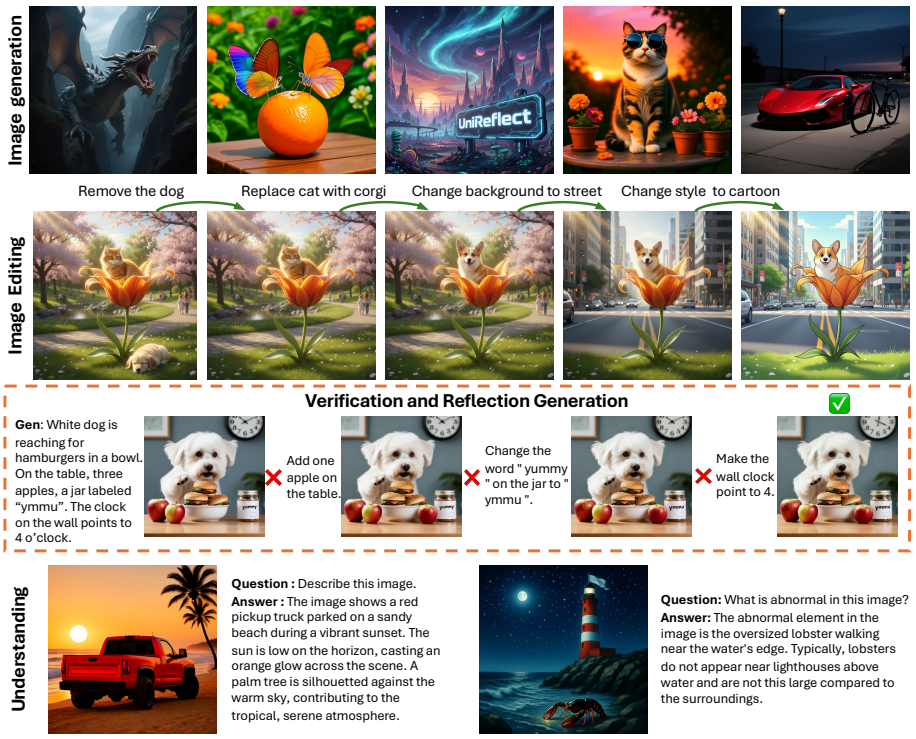

Fig. 1: Overview of the versatile capabilities of UniReflect. As a unified framework for visual understanding and generation, UniReflect excels in four key areas: (1) Image Generation: synthesizing high-fidelity images from complex prompts; (2) Image Editing: performing precise, multi-turn editing (e.g., style transfer and object replacement); (3) Verification and Reflection: uniquely performing in-model verification to detect semantic mismatches (e.g., missing objects or incorrect text) and generating structured reflections to guide Test-Time Scaling (TTS) for image refinement without external verifiers; and (4) Visual Understanding: engaging in detailed image captioning and reasoning.

Despite recent advances, the majority of existing methods conceptualize verification as an auxiliary, post-generation process. For instance, many contemporary pipelines [62, 74, 78] employ Visual Large Language Models (VLLMs) to assess the semantic consistency between a given prompt and its corresponding generated image. Although such approaches can be effective, their dependence on lexical feedback introduces notable computational overhead and latency. Moreover, this separation of verification from the core generation process results in fragmented information flows, thereby hindering the development of fully integrated multi-modal reasoning systems.

Motivated by this perspective, this study focuses on the central research question: **How can in-model verification be effectively implemented for multi-modal generative models?** Given that verification inherently involves visual understanding, such models should ideally integrate both understand-

ing and generation within a unified framework. Recent advancements in unified modeling approaches [6, 26, 34, 59, 62, 64] have established robust baselines for general-purpose visual understanding and content generation, thereby providing a solid foundation for pursuing deeper integration of verfication and reflection.

In this study, we introduce UniReflect, a unified framework for visual understanding and generation that incorporates in-model verification through a novel self-reflection tuning mechanism. Central to our approach is the integration of a Semantic Perceiver module, which establishes a bidirectional link between the understanding and generation processes. We further introduce a specialized token to construct a dedicated embedding space, enabling precise guidance during the decoding of refined images. The two-stage training recipe leverages a curated dataset encompassing verification, reflection, understanding, and generation tasks, thereby instructing UniReflect to: (1) evaluate the semantic correspondence between the input prompt and the generated image; (2) produce a structured analysis identifying mismatches or omitted elements; and (3) generate a refined image in a single inference pass. This design facilitates seamless test-time scalability entirely within the model, obviating the need for external verification modules or supplementary editing prompts.

Extensive experimental evaluations indicate that UniReflect attains state-of-the-art performance in image generation fidelity, while simultaneously enabling robust visual verification capabilities. By integrating visual comprehension with generative synthesis within a unified framework, our method establishes a novel paradigm for multi-modal modeling. This advancement facilitates the development of more efficient, self-adaptive generative systems capable of real-time reasoning about their outputs. Collectively, the main contributions of this paper can be summarized as:

· We propose UniReflect, a unified architecture that seamlessly integrates visual understanding and high-fidelity image generation. The design incorporates a Semantic Perceiver module and a specialized embedding space for the guidance of refined image, enabling bidirectional interaction between comprehension and synthesis for in-model semantic verification and reflection, thereby eliminating the need for external verification modules.

· We introduce a two-stage training recipe on a curated multi-task dataset that jointly covers verification, reflection, understanding, and generation. This strategy explicitly instructs the model to assess semantic alignment between prompts and generated images, produce structured analyses of mismatches or omissions, and regenerate refined images in a single inference pass.

· Our UniReflect model achieves state-of-the-art performance in image generation fidelity while providing robust verification capabilities. The results validate the effectiveness of integrating reasoning and generation within a single model, offering a new paradigm for efficient, self-adaptive multi-modal generative systems.

2 Related Works

Text-to-Image Generation Models. Diffusion models [22] become the dominant paradigm for image generation by iteratively denoising from noise to structured outputs, initially relying on U-Net architectures. Diffusion Transformers (DiT) [46] replace convolutions with global attention, enabling better modeling of complex spatial dependencies. Recent works such as PixArt-Σ [7] and Scale-DiT [75] scale generation to ultra-high resolutions, while DiT4Edit [16] adapts DiT for efficient image editing without fine-tuning. Efficiency-oriented variants like Dynamic Diffusion Transformer (D^{2}iT) [25] reduce inference cost, and multimodal extensions (MM-DiT) [72] enable controllable layout generation, highlighting DiT's versatility in modern generative modeling.

Multimodal Understanding Models. The rapid evolution of Large Language Models (LLMs) has significantly catalyzed the development of Visual Large Language Models (VLLMs), enabling robust vision-language co-understanding [2, 3, 30, 35, 38, 77]. Prominent examples such as BLIP-2 [30], Flamingo [2], MiniGPT-4 [77], LLaVA [38], InstructBLIP [14], and Qwen-VL [3] have demonstrated impressive performance in general visual reasoning. To transcend simple global descriptions and achieve fine-grained perception, subsequent works have integrated detailed grounding capabilities. Methods like Shikra [10], Ferret [69], Kosmos-2 [47], and VisionLLM [58] utilize bounding box regression, while LISA [28], PixelLM [52], GLaMM [51], and PerceptionGPT [48] employ mask decoders for precise segmentation. Although recent attempts like PSALM [76] seek to further unify these perception tasks, they remain fundamentally limited to the scope of understanding. Crucially, despite their advanced reasoning and localization abilities, these architectures are restricted to outputting text or coordinates and intrinsically lack the capability to perform visual generation or pixel-level image synthesis.

Unified understanding and generation model. There a re two design philosophies regarding unifying multimodal understanding and generation. One approach involves integrating understanding and generation into a single LLM, thereby unifying the training and optimization methods for both processes, commonly referred to as *native unified models*. Chameleon [56] and Show-o [65] integrate understanding and generation by integrating autoregressive text generation with discrete or continuous diffusion for images within unified model framework. Emu3 [59] and Unitoken [26] further focuses on enhancing the semantics of discrete tokens that constitute the unified visual representation. Another strategy is the *assembled unified models* approach, which leverages the concept of adapter modules by employing tuning adapters, learnable tokens, or hidden states as bridges. This methodology effectively connects these models with specialized VLLM and visual generative models such as Diffusion Transformers. OmniGen2 [62] leverages the hidden states of multimodal interleaved conditions produced by the VLLM, along with low-level visual features encoded by the VAE, as inputs for the visual generation model. BLIP3-o [6] and Uniworld [34] further employs visual features extracted by semantic encoders rather than VAEs generated image quality. These approach effectively preserves the

robust visual understanding capabilities of the original MLLM while facilitating the training and optimization of generative models.

Verfication and reflection for visual generation. Benefiting from the multimodal reasoning capabilities of vision-language models, the verifier offers essential reflection and refinement capabilities on visual outcomes during the reasoning and generation processes. OmniGen2 [62] leveraging a general VLLM as the verification and reflection model to construct reflection data for improving subsequent visual generation models. ReflectionFlow [78] and OmniVerifier [74] employ a specially trained robust VLLM as a verifier to rigorously assess the alignment of image content with input instructions. Subsequently, they provide detailed suggestions for modification to address any identified inconsistencies. These specially constructed data and the pipeline continuously improve the quality of subsequent generation.

3 Method

3.1 Overview

Overall architecture. The architecture of UniReflect is illustrated in Fig. 2, which consists of a frozen VLLM, a semantic perceiver, a VAE, and a diffusion transformer to generate visual outcomes for text-to-image and image editing tasks according to the user's instruction. The VLLM performs multiple general multimodal understanding tasks while the verifier head executes verification and sequential reflection for the visual generation. The proposed Semantic Perceiver module integrates verification details and semantic instruction information for visual generation leveraging shared weights, which facilitates the synergy between verification and generation tasks through multi-task joint training. The VLLM takes three types of inputs according to the different tasks: visual-text pairs for multimodal understanding, reflection instructions and image to be verified for visual generation verification, generation instructions or image edit pairs. The output embeddings of the <verifier> token, <reflect> token, and semantic instruction <img> token for visual generation are further fed into the verifier head and diffusion transformer after the fusion of semantic perceiver. Besides, we utilize the VAE to efficiently encode the low-level visual features for diffusion processing.

Visual Large Language Model. We take a pre-trained VLLM as our powerful multi-modal information processor for understanding, verification, and generation tasks, which not only preserves superior understanding capabilities but also integrates verifier details and high-level generation semantics from multiple instructions and visual input.

We utilize a unified form to represent aforementioned three tasks. Specifically, the model takes vision-text pairs $\{(\mathcal{V}, \mathcal{T})\}$ as inputs, where $\mathcal{V}$ and $\mathcal{T}$ are transformed into separate tokens through visual and textual tokenizers to facilitate the comprehensive understanding of multi-modal inputs through the fusion process of VLLM F_{vllm}. Formally,

$$E = F_{vllm}(\mathcal{V}, \mathcal{T}), \tag{1}$$

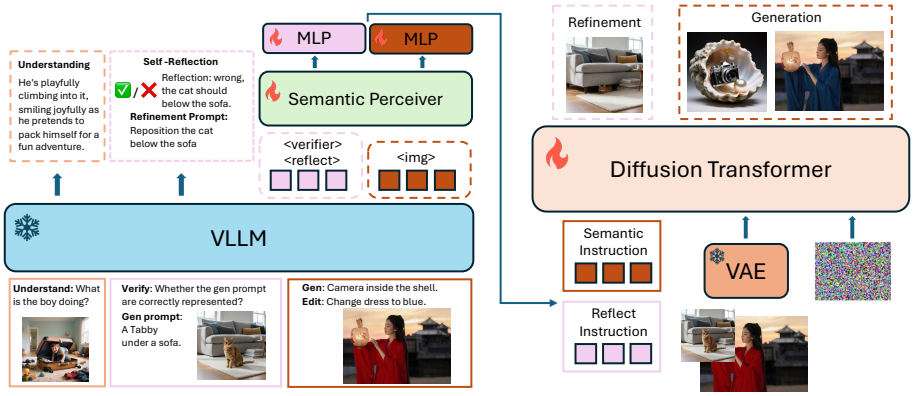

Fig. 2: Overview of UniReflect. We present a unified framework that seamlessly integrates visual understanding, verification, and generation. The architecture comprises a frozen Visual Large Language Model (VLLM) for multimodal reasoning and a Diffusion Transformer (DiT) for high-fidelity image generation. (1) The Semantic Perceiver: The core innovation is the Semantic Perceiver module, which acts as a bidirectional bridge between the VLLM and the DiT. It processes task-specific tokens like <verifier>, <reflect>, and <img> which are extracted from the VLLM's hidden states. (2) Unified Capabilities: This design enables three distinct modes within a single model: Visual Understanding, Verification (assessing prompt-image alignment), and Generation/Editing (synthesizing images based on semantic and reflection instructions). (3) Decoupled Verification: Unlike standard VLLMs, our verifier head separates scoring from explanation, producing a verification score to trigger necessary refinements.

where E denotes the output embeddings of VLLM. E are passed to the LM head to generate textual output for understanding tasks. Furthermore, we manually extract the <verifier> token, <reflect> token, and semantic instruction <img> token from E and feed them into semantic perceiver to obtain $E_{\mathcal{V}}$, $E_{\mathcal{R}}$ and $E_{\mathcal{S}}$, which are further fed into the verifier head and diffusion decoder separately to generate verification scores, reflection result and visual generation respectively.

Visual generation. Diffusion decoder F_D generates the images Y through the similar flow matching process [15, 34, 62] of four inputs: the semantic instruction embedding $E_{\mathcal{S}}$, the reflection instruction embedding $E_{\mathcal{R}}$, the original VAE latents X_v for image editing task, and the diffusion noise N_o. Formally,

$$Y = F_D(E_{\mathcal{S}}, E_{\mathcal{R}}, X_v, N_o), \tag{2}$$

The architecture of diffusion decoder F_D is a simple diffusion transformer which concatenates all the features aforementioned, facilitating joint learning from different modalities. Besides, we adopt classifier-free guidance during the visual content generation.

Training objectives. The model can be trained jointly on multi-task datasets using the unified loss $\mathcal{L}$. Specifically, we employ an autoregressive cross-entropy

loss $\mathcal{L}_{text}$ for text prediction, a cross-entropy loss $\mathcal{L}_{cls}$ for verification classification, and a MSE Loss $\mathcal{L}_{visual}$ for visual reconstruction. λ indicates the sum weight. Formally,

$$\mathcal{L} = \mathcal{L}_{text} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{visual}\mathcal{L}_{visual}, \tag{3}$$

3.2 Semantic Perceiver

Why integration of verification and generation? Previous unified multimodal models [34, 45, 62] for understanding and generation often employ frozen VLLMs as the text prompt encoder for the instructions of visual generation. Though impressive, they may suffer suboptimal generation results since the VLLMs are pre-trained for question-answering which may lead to gap between understanding and precise generation. In our work, we leverage a shared Semantic Perceiver to unify the semantic space, enabling the model to better capture high-level semantics and fine-grained generative details during the verification and reflection learning, thereby facilitating more accurate and effective generation. Notably, the verification process and visual generation are intrinsically complementary, enjoying the benefits of multi-task learning.

Specifically, we input the <verifier> token, <reflect> token, and semantic instruction <img> token extracted from the final hidden states of VLLM as input to the semantic perceiver. The semantic perceiver, implemented as a multi-layer transformer block, efficiently integrates multimodal information across distinct tasks. This integration facilitates complementary interactions between verification and generation processes, thereby enhancing overall model performance.

Decoupling verification strategy. Previous verifiers [74] for judging the quality and accuracy of image generation tend to directly output the True or False text for verifying classification along with optional explanation, which suffer limited verification performance due to the unstable abilities to follow instructions. In UniReflect, we utilize a decoupling verification strategy to generate verification scores simply employ a class head on <verifier> token. Furthermore, we obtain explicit explanation and reflection text through the next token prediction on the condition of <verifier> token if the verification score is False.

The refinement of the generated image is completed upon the output reflect embedding $E_{\mathcal{R}}$ of semantic perceiver while previous TTS methods [74] need to input the reflection and editing prompt into the additional image editing model which is complex and inefficient.

3.3 Test-Time Scaling via Sequential Reflection

Our proposed unified multimodal framework, UniReflect, fundamentally alters the paradigm of generative refinement by integrating text-to-image synthesis, visual verification, and image editing within a single, cohesive model architecture. Unlike existing approaches that rely on disjointed pipelines involving external Visual Large Language Models (VLLMs) for feedback, UniReflect possesses intrinsic advantages in generation, reflection, and refinement. This unity allows for

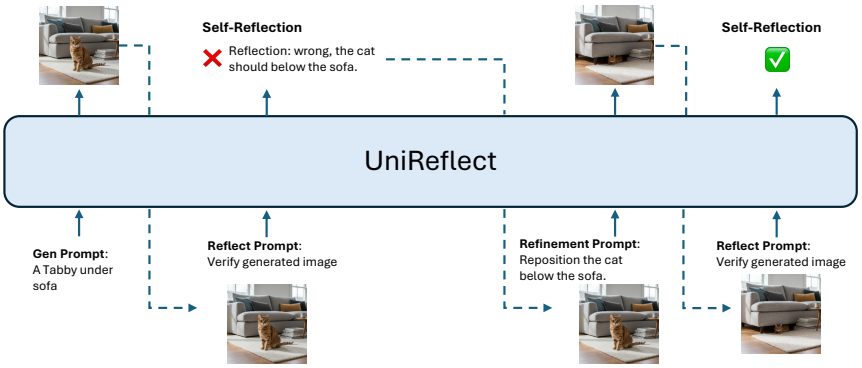

Fig. 3: Illustration of the Sequential Test-Time Scaling (TTS) Pipeline via Self-Reflection. This diagram demonstrates UniReflect's capability to iteratively refine generation quality without external verifiers.(1) Generation and Verification: Given an initial prompt, the model generates an initial image. The unified verifier immediately assesses semantic alignment.(2) Self-Reflection: Upon detecting a mismatch (e.g., the cat is on rather than under the sofa), the model generates a structured textual reflection explaining the error.(3) Refinement: This reflection is encoded into a reflection embedding ($E_{\mathcal{R}}$) via the Semantic Perceiver, which guides the Diffusion Transformer to regenerate the image.(4) Closed-Loop Optimization: This process forms an autonomous "generation-verification-correction" loop, significantly enhancing complex instruction following and spatial reasoning during inference.

a seamless flow of semantic information, obviating the computational overhead and information loss associated with cascading distinct models.

As illustrated in Fig. 3, we introduce UniReflect-TTS, a sequential Test-Time Scaling (TTS) pipeline designed to iteratively enhance generation fidelity through self-supervised correction. The pipeline operates as an autonomous closed-loop system during inference:

1. **Generation and Verification:** Given an initial input prompt, the model synthesizes a candidate image. The unified verifier head then immediately assesses the semantic alignment between the prompt and the generated visual content.
2. **Structured Self-Reflection:** If the verification score falls below a confidence threshold or indicates a semantic mismatch (e.g., missing objects or incorrect spatial relations), the model triggers its reflection mechanism. Instead of merely outputting a binary label, the model leverages the Decoupling Verification strategy to generate a detailed textual analysis explaining the discrepancy.
3. **Refinement via Semantic Embedding:** This textual reflection is not used as a raw prompt; rather, it is processed by the Semantic Perceiver to produce a specialized reflection embedding, denoted as $E_{\mathcal{R}}$. This embedding encapsulates high-level semantic guidance specifically tuned to correct the identified errors.

The refinement process is mathematically formulated as a conditional generation task where the updated image Y_{t+1} is synthesized by the diffusion decoder F_D. Crucially, the decoder is conditioned on both the original semantic instruction $E_{\mathcal{S}}$ and the newly generated reflection embedding $E_{\mathcal{R}}$. This allows UniReflect to preserve the correct elements of the original generation while precisely rectifying the erroneous regions guided by its own criticism. By cycling through this verify-reflect-refine loop, UniReflect achieves robust test-time scaling, effectively converting compute time into generation quality without the need for human intervention or external gradient guidance.

4 Experiment

Datasets. Our training corpus is constructed from three primary tasks: text-to-image (T2I) generation, image editing, and visual verification. For the T2I generation task, we employ a combination of publicly available datasets—namely BLIP3-o [6], LAION-Aesthetic [54], JourneyDB [55], and Open-Sora Plan [33]—supplemented by 0.6 million internally curated samples. This results in a total of 3.9 million instances for image generation. For the image editing task, we compile 1 million samples from ImgEdit [68], SEED-X [17], OmniEdit [60], and ShareGPT-4o [9]. Finally, for the visual verification task, we adopt the data curation pipeline outlined in OmniVerifier [74], yielding 20,000 high-quality verification samples.
Implementation Detail. The proposed UniReflect framework integrates Qwen3-VL as the MLLM and employs a single-stream DiT architecture for image synthesis. These components are connected via the proposed semantic perceiver. To retain the MLLM's inherent multimodal comprehension capabilities, the majority of its parameters remain frozen during training; only the newly introduced special tokens (<verifier>, <reflect> and <img>) are updated. The DiT backbone is initialized from OmniGen2 [62], whereas the semantic perceiver is trained from scratch. Training is conducted in two stages: (1) understanding–generation alignment, leveraging datasets for both image generation and image editing; and (2) joint learning across text-to-image generation, image editing, and visual verification tasks, thereby enhancing the model's overall capacity. The complete set of hyperparameters and training configurations is provided in Tab. 1.

Table 1: Hyperparameters for both training stages.

Parameters	Stage1	Stage2
Training Components	Semantic Perceiver + DiT	
Optimizer	AdamW	
Training Rate	1×10^{-3}	1×10^{-6}
Batch Size	128	64
Number of Steps	40000	30000
Learning Rate Schedule	Cosine Decay	
Weight Decay	0.0	0.05
Warmup Ratio	0.03	
Training Task	Gen. Edit.	Gen. Edit. Veri.
Image Size	512×512	1024×1024

Table 2: Overall comparison across Understanding, Image Generation, Image Editing, and Visual Verification tasks (model-based evaluation). *: The first term represents the number of parameters of MLLM, while the second term corresponds to the size of the image generation head. † refers to the methods using prompt engineering.

Method	# Params	Understanding			Image Generation		Image Editing		Visual Verification
		MMBench	MMMU	MathVista	GenEval	DPG	ImgEdit	GEdit-EN	ViVerBench
Understanding									
LLaVA-v1.5 [36]	13B	36.4	67.8	27.6	-	-	-	-	-
LLaVA-NexT [37]	13B	79.3	51.1	46.5	-	-	-	-	-
InternVL2.5 [12]	7B	84.6	56.0	64.4	-	-	-	-	-
Image Generation									
SDXL [49]	-	-	-	-	0.55	74.7	-	-	-
SD3-medium [1]	-	-	-	-	0.62	84.1	-	-	-
FLUX.1-dev [27]	-	-	-	-	0.66	84.0	-	-	-
Image Editing									
Instruct-P2P [5]	-	-	-	-	-	-	1.88	3.68	-
MagicBrush [73]	-	-	-	-	-	-	1.90	1.86	-
AnyEdit [70]	-	-	-	-	-	-	2.45	3.21	-
Step1X-Edit [39]	-	-	-	-	-	-	3.06	**6.70**	-
Visual Verification									
Qwen 2.5-VL 7B [4]	7B	83.5	58.6	68.2	-	-	-	-	0.523
OmniVerifier [74]	7B	-	-	-	-	-	-	-	0.559
Understanding and Generation									
Show-o [65]	1.3B	-	27.4	-	0.68	67.3	-	-	-
Janus-Pro [11]	1.5B	75.5	36.3	-	0.80	84.2	-	-	-
Emu3 [59]	8B	58.5	31.6	-	0.54 / 0.66†	80.6	-	-	-
MetaQuery [45]	7B + 1.6B*	83.5	58.6	68.2	0.80†	82.1	-	-	-
BLIP3-o [6]	7B + 1.4B*	83.5	50.6	-	0.84†	81.6	-	-	-
BAGEL [15]	7B + 7B*	85.0	55.3	73.1	0.82 / **0.88**†	85.1	3.20	6.52	-
UniWorld-V1 [34]	7B + 12B*	83.5	58.6	68.2	0.84†	81.4	3.26	4.85	-
OmniGen2 [62]	3B + 4B*	79.1	53.1	62.3	0.80 / 0.86†	83.6	**3.44**	6.42	-
UniReflect	4B + 4B*	**85.1**	**67.4**	**73.7**	**0.83** / 0.87†	**86.7**	3.28	5.14	**0.579**

Benchmarks. We conduct a comprehensive evaluation across multiple benchmarks to assess the unified capabilities of our model in a broad range of generative tasks. The overall comparison with state-of-the-art approaches is summarized in Tab. 2. For visual understanding, we adopt MMBench [40], MMMU [71], and MathVista [42] as validation benchmarks. Compositional reasoning is evaluated using GenEval [18], while long-prompt following in text-to-image (T2I) generation is assessed through DPG-Bench [23]. Furthermore, UniReflect is tested on two image editing benchmarks—GEdit-Bench-EN [39] and ImgEdit-Bench [68]—to examine editing proficiency. Finally, the model's visual verification capabilities are evaluated using the standard benchmark ViVerBench [74] with the model-based method.

4.1 Main Results

Visual Understanding. Given that the MLLM component remains frozen during training, UniReflect is anticipated to preserve visual understanding capabilities comparable to those of Qwen3-VL-4B-instruct. As illustrated in Tab. 2, this architectural choice allows our approach to attain competitive performance relative to recent state-of-the-art models, while requiring substantially less training data and computational overhead.

Table 3: Evaluation of text-to-image generation ability on GenEval [18] benchmark. † refers to the methods using prompt engineering. TTS means Test-Time Scaling for image generation refinement.

Method	Single object↑	Two object↑	Counting↑	Colors↑	Position↑	Color attribution↑	Overall↑
			Gen. Only				
SDv2.1 [53]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
SDXL [49]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
SD3-medium [1]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev [27]	0.99	0.81	0.79	0.74	0.20	0.47	0.67
OmniGen [63]	0.98	0.84	0.66	0.74	0.40	0.43	0.68
			Unified				
TokenFlow-XL [50]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus [61]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Janus Pro [11]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Emu3† [59]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Show-o [65]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
MetaQuery† [45]	-	-	-	-	-	-	0.80
BLIP3-o† [6]	-	-	-	-	-	-	0.84
BAGEL [15]	0.99	0.94	0.81	0.88	0.64	0.63	0.82
BAGEL† [15]	0.98	0.95	0.84	0.95	0.78	0.77	<u>0.88</u>
UniWorld-V1 [34]	0.99	0.93	0.79	0.89	0.49	0.70	0.80
UniWorld-V1† [34]	0.98	0.93	0.81	0.89	0.74	0.71	0.84
OmniGen2 [62]	1.00	0.95	0.64	0.88	0.55	0.76	0.80
OmniGen2† [62]	0.99	0.96	0.74	0.98	0.71	0.75	0.86
UniReflect	1.00	0.99	0.65	0.98	0.63	0.72	0.83
UniReflect-OmniVerifier-TTS	1.00	0.99	0.73	0.98	0.71	0.75	0.86
UniReflect-TTS	1.00	1.00	0.77	0.98	0.76	0.81	0.89

Text-to-Image Generation. The comparative results across GenEval and DPG-Bench collectively demonstrate the effectiveness of UniReflect in handling complex, compositional prompts and generating high-quality visual outputs. On the GenEval benchmark (Tab. 3), UniReflect attains a score of 0.87, closely matching the state-of-the-art BAGEL model's performance (0.88). This achievement is particularly notable given the substantial disparity in model scale and training data: BAGEL employs 14B trainable parameters and is trained on 1.6B image–text pairs, whereas UniReflect reaches comparable accuracy with only 4B trainable parameters and a significantly smaller corpus of 4M image–text pairs. Further enhancement is obtained through a test-time scaling mechanism, which raises the GenEval score to 0.89 and yields marked improvements in sub-tasks such as object counting, spatial localization, and color attribution.

The strength of UniReflect is further corroborated by its performance on DPG-Bench (Tab. 4), where it achieves the highest scores across all evaluated dimensions, culminating in an overall score of 86.69. This surpasses both BAGEL (85.07) and the leading specialized model, SD3-medium (86.69). The convergence of superior results across two distinct evaluation frameworks provides compelling empirical evidence of UniReflect 's capacity to accurately interpret intricate textual instructions and to synthesize visually coherent and semantically faithful images, even under constrained model and data budgets.

Image Editing. As presented in Tab. 5, the proposed UniReflect achieves performance on par with state-of-the-art open-source counterparts, attaining an overall score of 3.28, which is only marginally lower than that of OmniGen2 (3.44). To further evaluate its out-of-domain generalization capability, we report results on the GEdit dataset in Tab. 6. The observed performance degradation

Table 4: Evaluation of text-to-image generation ability on DPG-Bench [23] benchmark.

Method	Global↑	Entity↑	Attribute↑	Relation↑	Other↑	Overall↑
			Gen. Only			
SDXL [49]	83.27	82.43	80.91	86.76	80.41	74.65
PlayGroundv2.5 [29]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [31]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ [8]	86.89	82.89	88.94	86.59	87.68	80.54
DALLE3 [43]	90.97	89.61	88.39	90.58	89.83	83.50
SD3-medium [1]	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev [27]	82.1	89.5	88.7	91.1	89.4	84.0
OmniGen [63]	87.90	88.97	88.47	87.95	83.56	81.16
			Unified			
Show-o [65]	79.33	75.44	78.02	84.45	60.80	67.27
EMU3 [59]	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL [50]	78.72	79.22	81.29	85.22	71.20	73.38
Janus [61]	82.33	87.38	87.70	85.46	86.41	79.68
Janus Pro [11]	86.90	88.90	89.40	89.32	89.48	84.19
BLIP3-o 4B [6]	-	-	-	-	-	79.36
BLIP3-o 8B [6]	-	-	-	-	-	81.60
BAGEL [15]	88.94	90.37	91.29	90.82	88.67	85.07
UniWorld-V1 [34]	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2 [62]	88.81	88.83	90.18	89.37	90.27	83.57
UniReflect	**92.78**	**91.03**	**91.54**	**93.24**	**90.34**	**86.69**

Table 5: Comparison results on ImgEdit-Bench [68]. "Overall" is calculated by averaging all scores across tasks.

Method	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall ↑
GPT-4o [44]	**4.61**	**4.33**	**2.9**	**4.35**	**3.66**	**4.57**	**4.93**	**3.96**	**4.89**	**4.2**
MagicBrush [73]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-P2P [5]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.2	1.46	1.88
AnyEdit [70]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
OmniGen [63]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
Step1X-Edit [39]	**3.88**	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
BAGEL [15]	3.56	3.31	1.7	3.3	2.62	3.24	4.49	2.38	4.17	3.2
UniWorld-V1 [34]	3.82	**3.64**	**2.27**	3.47	**3.24**	2.99	4.21	**2.96**	2.74	3.26
OmniGen2 [62]	3.57	3.06	1.77	**3.74**	3.2	**3.57**	**4.81**	2.52	**4.68**	**3.44**
UniReflect	3.36	3.04	2.02	3.25	2.98	3.51	4.68	2.7	3.99	3.28

can be primarily attributed to the limited quantity and diversity of available editing data, as well as the relatively small proportion of such data in the overall training corpus—particularly due to the inclusion of the verification task. Nevertheless, the proposed test-time scaling strategy consistently yields performance gains under these conditions, improving the score from 5.14 to 6.07.

Verification and Reflection. As illustrated in Tab. 5 and Tab. 6, the proposed unified model—operating without reliance on any external verifier—exhibits reflective capabilities and yields notable performance improvements. Specifically, when applying Test-Time Scaling, the model's scores increase from 0.83 to 0.89 on GenEval and from 5.14 to 6.07 on GEdit. Furthermore, we extend our evaluation to the more challenging GenEval++ benchmark (Tab. 7) to assess compositional generation capabilities. UniReflect achieves a remarkable overall score of 0.654, significantly outperforming state-of-the-art baselines such as FLUX.1-dev

Table 6: Quantitative comparison GEdit-Bench-EN [39]. SC (Semantic Consistency) evaluates instruction following, and PQ (Perceptual Quality) assesses image naturalness and artifacts. Higher scores are better for all metrics.

Method	SC↑	PQ↑	O ↑
Gemini-2.0-flash [19]	6.73	6.61	6.32
GPT-4o [44]	7.85	7.62	7.53
Instruct-Pix2Pix [5]	3.58	5.49	3.68
MagicBrush [73]	4.68	5.66	4.52
AnyEdit [70]	3.18	5.82	3.21
OmniGen [63]	5.96	5.89	5.06
Step1X-Edit [39]	7.09	6.76	**6.70**
BAGEL [15]	**7.36**	<u>6.83</u>	<u>6.52</u>
UniWorld-V1 [34]	4.93	**7.43**	4.85
OmniGen2 [62]	<u>7.16</u>	6.77	6.41
UniReflect	5.34	5.93	5.14
UniReflect-TTS	6.14	6.81	6.07

Table 7: Evaluation of UniReflect and UniReflect-TTS on compositional generation benchmarks GenEval++ [67].

Method	Color Count	Color/Count	Color/Pos	Pos/Count	Pos/Size	Multi-Count	Overall
SD-3-Medium [1]	0.550 0.500	0.125	0.350	0.175	0.150	0.225	0.296
FLUX.1-dev [27]	0.350 0.625	0.150	0.275	0.200	0.375	0.225	0.314
Janus-Pro [11]	0.450 0.300	0.125	0.300	0.075	0.350	0.125	0.246
Bagel [15]	0.325 0.600	0.250	0.325	0.250	0.475	0.375	0.371
UniReflect	0.610 0.785	0.715	0.543	0.480	0.705	0.745	0.654
UniReflect-TTS	0.672 0.798	0.720	0.560	0.505	0.733	0.769	0.680

(0.314) and Bagel (0.371). This substantial margin highlights the efficacy of our unified architecture in resolving complex semantic alignments and intricate spatial relationships. Additionally, the results reported in Tab. 8 demonstrate that our 4B model, when trained exclusively on synthetic data, attains performance comparable to that of the customized 7B verification model (0.579 vs. 0.559). In contrast to the RL–based training employed in Omniverifier [74], our findings indicate that even a straightforward fine-tuning procedure is sufficient to surpass the baseline.

4.2 Ablations

Effect of unified verifier. As presented in Tab. 3, we evaluate the performance of different verifiers within the TTS framework. While our model exhibits marginally lower scores than OmniVerifier [74] in domains closely associated with image generation—such as *concept existence* and *object relation*—it consistently achieves superior results across all sub-tasks in GenEval using TTS. These findings underscore the effectiveness of the proposed unified training strategy in enhancing overall reflection performance.

Effect of the proposed design. We evaluate the effectiveness of the proposed Semantic Perceiver and Decoupling Verification strategy, as presented in Tab. 9. Incorporating the Semantic Perceiver yields substantial improvements in image

Table 8: Model-based evaluation of advanced VLMs on **ViVerBench**.

Method	Concept Existence			Object Relation		World Dynamics		Image Annotation			State Value Evaluation				STEM		Overall
	Obj.	Attr.	Abs.P.	Spat.	N-Spat.	S-Phy	D-Phy	BBox	Point	Count	Maze	F.Lake	Robot.	GUI	Chart	LaTeX	
GPT-4o [44]	0.509	0.603	0.630	0.525	0.642	0.658	0.279	0.581	0.670	0.445	0.490	0.643	0.513	0.727	0.596	0.737	0.578
OpenAI o1 [24]	0.670	0.754	0.692	0.704	0.739	0.658	0.593	0.633	0.704	0.478	**0.537**	0.646	0.563	0.718	0.712	0.897	0.669
OpenAI o4-mini	0.732	0.741	0.637	0.762	0.709	0.575	0.496	0.718	0.667	0.462	0.490	0.650	0.633	0.750	0.664	0.876	0.660
Seed 1.5-VL [21]	0.732	**0.763**	0.616	0.779	**0.799**	0.546	0.425	**0.839**	**0.807**	0.527	0.463	0.718	**0.671**	0.810	0.680	0.907	**0.693**
OpenAI o3	0.705	0.724	0.726	0.713	0.716	0.696	**0.614**	0.718	0.767	0.505	0.360	0.671	0.627	0.810	0.720	0.876	0.684
GPT-5	0.688	0.733	0.767	0.725	0.701	**0.721**	0.600	0.734	0.752	0.505	0.317	0.743	0.582	0.810	**0.744**	0.866	0.687
Gemini 2.5 Pro [13]	**0.746**	0.746	**0.801**	**0.875**	0.701	0.679	0.350	0.742	0.752	**0.582**	0.533	**0.804**	0.544	**0.843**	0.460	0.799	0.685
Qwen 2.5-VL 72B [4]	0.696	0.642	0.616	0.542	0.754	0.550	0.304	0.762	0.656	0.451	0.237	0.507	0.494	0.741	0.604	**0.922**	0.592
InternVL3.5 A28B [4]	0.683	0.737	0.589	0.729	0.746	0.550	0.267	0.730	0.696	0.440	0.260	0.532	0.462	0.722	0.604	0.876	0.601
Qwen 2.5-VL 7B [4]	0.526	0.547	0.493	0.500	0.604	0.487	0.307	0.601	0.548	0.440	0.453	0.404	0.627	0.588	0.536	0.711	0.523
Qwen 3-VL 4B [57]	0.602	0.672	0.568	0.504	0.597	0.483	0.117	0.725	0.581	0.434	0.176	0.482	0.474	0.736	0.452	0.798	0.525
Random	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
OmniVerifier 7B [74]	0.688	0.694	0.514	0.717	0.664	0.442	0.382	0.625	0.578	0.462	0.497	0.407	0.424	0.616	0.432	0.799	0.559
UniReflect 4B	0.679	0.591	0.493	0.604	0.597	0.667	0.214	0.569	0.407	0.527	0.637	0.643	0.703	0.565	0.564	0.799	0.579

Table 9: Effect of the proposed Semantic Perceiver and Decoupling verification strategy.

Semantic Perceiver	Decoupling	GenEval	ViVerBench
		0.77	0.525
✓		0.81	0.534
	✓	0.78	0.569
✓	✓	**0.83**	**0.579**

generation performance, whereas the Decoupling Verification strategy enhances visual verification accuracy. When combined, the proposed UniReflect framework demonstrates consistent gains across both tasks relative to the baseline, achieving an improvement of +0.050 on GenEval and +0.054 on ViVerBench.

5 Conclusion

In this work, we presented UniReflect, the first unified framework that seamlessly integrates visual understanding, generation, and in-model verification through self-reflection tuning. Unlike prior approaches that rely on external Visual Large Language Models for prompt-image alignment checks, UniReflect directly performs semantic verification within the generation process, producing structured analyses and refinement cues that enable efficient test-time scaling without additional modules or prompts. Our experiments confirm that this unified approach not only improves image fidelity but also delivers consistent and reliable visual verification, setting a new performance benchmark for multi-modal methods. We believe that the principles demonstrated here—tight coupling of comprehension and synthesis, and in-model reflection—open promising directions for future research in scalable, efficient, and trustworthy multi-modal AI systems.

References

1. AI, S.: Sd3-medium. https://stability.ai/news/stable-diffusion-3-medium (2024)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
6. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
7. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: European Conference on Computer Vision. pp. 74–91. Springer (2024)
8. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
9. Chen, J., Cai, Z., Chen, P., Chen, S., Ji, K., Wang, X., Yang, Y., Wang, B.: Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. arXiv preprint arXiv:2506.18095 (2025)
10. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
11. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
12. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023), https://arxiv.org/abs/2305.06500
15. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)

16. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: Dit4edit: Diffusion transformer for image editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2969–2977 (2025)
17. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)
18. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36** (2024)
19. Google: Gemini 2.0 flash. `https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation` (2025)
20. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
21. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
23. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
24. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
25. Jia, W., Huang, M., Chen, N., Zhang, L., Mao, Z.: Dˆ 2it: Dynamic diffusion transformer for accurate image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12860–12870 (2025)
26. Jiao, Y., Qiu, H., Jie, Z., Chen, S., Chen, J., Ma, L., Jiang, Y.G.: Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3600–3610 (2025)
27. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
28. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. ArXiv **abs/2308.00692** (2023), `https://api.semanticscholar.org/CorpusID:260351258`
29. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024)
30. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
31. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024)
32. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.: Let's verify step by step. In: The Twelfth International Conference on Learning Representations (2023)
33. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)

34. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
35. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
36. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
37. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
38. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
39. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
40. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
41. Liu, Z., Wang, P., Xu, R., Ma, S., Ruan, C., Li, P., Liu, Y., Wu, Y.: Inference-time scaling for generalist reward modeling. arXiv preprint arXiv:2504.02495 (2025)
42. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
43. OpenAI: Dall·e 3. https://openai.com/index/dall-e-3 (2024)
44. OpenAI: Gpt-4o. https://openai.com/index/introducing-4o-image-generation (2025)
45. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
46. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
47. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
48. Pi, R., Yao, L., Gao, J., Zhang, J., Zhang, T.: Perceptiongpt: Effectively fusing visual perception into llm. arXiv preprint arXiv:2311.06612 (2023)
49. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
50. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2545–2555 (2025)
51. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. arXiv preprint arXiv:2311.03356 (2023)
52. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. ArXiv **abs/2312.02228** (2023), https://api.semanticscholar.org/CorpusID:265659012

53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
54. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
55. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Advances in Neural Information Processing Systems **36** (2024)
56. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
57. Team, Q.: Qwen3 technical report (2025), `https://arxiv.org/abs/2505.09388`
58. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36** (2024)
59. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
60. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: Omniedit: Building image editing generalist models through specialist supervision. In: The Thirteenth International Conference on Learning Representations (2024)
61. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
62. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
63. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13294–13304 (2025)
64. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
65. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
66. Yang, Z., Chen, L., Cohan, A., Zhao, Y.: Table-r1: Inference-time scaling for table reasoning. arXiv preprint arXiv:2505.23621 (2025)
67. Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)
68. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
69. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023)

70. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
71. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: 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. pp. 9556–9567 (2024)
72. Zhang, H., Hong, D., Wang, Y., Shao, J., Wu, X., Wu, Z., Jiang, Y.G.: Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18487–18497 (2025)
73. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
74. Zhang, X., Zhang, X., Wu, Y., Cao, Y., Zhang, R., Chu, R., Yang, L., Yang, Y.: Generative universal verifier as multimodal meta-reasoner. arXiv preprint arXiv:2510.13804 (2025)
75. Zhang, Y., Tai, Y.W.: Scale-dit: Ultra-high-resolution image generation with hierarchical local attention. arXiv preprint arXiv:2510.16325 (2025)
76. Zhang, Z., Ma, Y., Zhang, E., Bai, X.: Psalm: Pixelwise segmentation with large multi-modal model. arXiv preprint arXiv:2403.14598 (2024)
77. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
78. Zhuo, L., Zhao, L., Paul, S., Liao, Y., Zhang, R., Xin, Y., Gao, P., Elhoseiny, M., Li, H.: From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15329–15339 (2025)