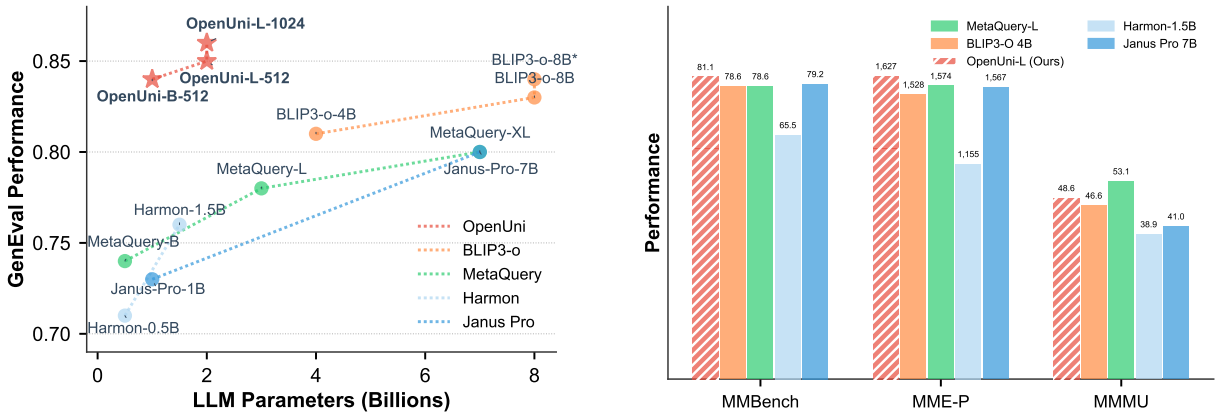
OpenUni: A Simple Baseline for Unified Multimodal Understanding and Generation

Size Wu[*1] **Zhonghua Wu**[*2] **Zerui Gong**[*1]
Qingyi Tao[2] **Sheng Jin**[3] **Qinyue Li**[2] **Wei Li**[1] **Chen Change Loy**[1]
[1] S-Lab, Nanyang Technological University [2] SenseTime Research [3] SenseTime Research and Tetras.AI
size001@e.ntu.edu.sg wei.l@ntu.edu.sg ccloy@ntu.edu.sg

ABSTRACT

In this report, we present *OpenUni*, a simple, lightweight, and fully open-source baseline for unifying multimodal understanding and generation. Inspired by prevailing practices in unified model learning, we adopt an efficient training strategy that minimizes the training complexity and overhead by bridging the off-the-shelf multimodal large language models (LLMs) and diffusion models through a set of learnable queries and a light-weight transformer-based connector. With a minimalist choice of architecture, we demonstrate that OpenUni can: 1) generate high-quality and instruction-aligned images, and 2) achieve exceptional performance on standard benchmarks such as GenEval, DPG-Bench, and WISE, with only 1.1B and 3.1B activated parameters. To support open research and community advancement, we release all model weights, training code, and our curated training datasets (including 23M image-text pairs) at `https://github.com/wusize/OpenUni`. [1]



(a) Evaluation of text-to-image generation: performance of OpenUni variants and baselines on GenEval versus parameter count.

(b) Multimodal understanding: comparison on MMBENCH, MME-P and MMMU.

Figure 1: OpenUni delivers strong performance with efficient parameter usage across both *generation* and *understanding* tasks.

1 Introduction

The landscape of multimodal artificial intelligence has been dominated by the recent progress of multimodal large language models (LLMs) [1, 2, 3, 4, 5, 6, 7, 8] and diffusion models [9, 10, 11, 12, 13, 14, 15, 16], driven by architectural innovations and computational scalings of transformers [17]. To further advance the frontier of multimodal

[*]Equal contribution.
[1]This is an ongoing project.

intelligence, a natural leap forward would be integrating the two minds of understanding and generation into a single brain, demonstrated by GPT4-o [18]'s impressive instruction-following ability of content generation.

Existing research efforts that unify multimodal understanding and generation can typically be divided into two subtracks. One type of work [19, 20, 21, 22, 23, 24, 25] explores native multimodal models from scratch and shares the parameters of LLM for both tasks. Another line of work stitches LLMs and generation models to build unified frameworks [26, 27, 28, 29, 30]. More recently, MetaQuery [31] and BLIP3-o [29] directly align frozen multimodal LLMs with diffusion models, effectively instilling generation ability into an already established multimodal system. These works unveil the potential of a simple connection module that transfers the knowledge of well-trained LLMs to generation models, for controllable and high-quality visual generation. Inspired by these studies, we present *OpenUni*, an open-source framework for unified multimodal understanding and generation, with minimum architectural complexities and computational overhead.

Specifically, OpenUni adheres to the simplest design choices presented in MetaQuery [31], using only learnable queries and a light-weight connector between a multimodal LLM (MLLM) and a diffusion model. A two-stage training recipe is adopted to build OpenUni. In the first (pre-training) stage, we align the LLM and diffusion model on 23M image-text pairs, by training only the learnable queries and the connector. The 23M training images used in this stage are sourced from public datasets and re-captioned by LLMs, which will also be released. In the finetuning stage, we unlock the diffusion model and train OpenUni on the 60k high-quality images contributed by BLIP3-o [29].

We implemented three model variants, namely OpenUni-B-512 and OpenUni-L-512 and OpenUni- L-1024, characterized by different model sizes and image resolutions. For image understanding, OpenUni inherits its base MLLM's strong performance on multimodal question-answering benchmarks. For image generation, our smaller variant OpenUni-B-512 achieves a score of 0.84 on GenEval, on par with BLIP3-o-8B [32], using only 1.1B activated parameters, while significantly outperforming prior unified models like Janus-Pro [23]. Meanwhile, OpenUni-L-1024 archives the best performance (0.86) among open-source unified models on GenEval with 3.1B activated parameters. Besides, OpenUni exhibits competitive performance on the WISE benchmark that assesses world knowledge comprehension, surpassing models that utilize LLMs of similar scales. To support research and reproducibility, we release the full framework as an open-source baseline with minimal training complexity, modular design, publicly available training data, and an out-of-the-box training pipeline.

2 Related Work

Multimodal LLMs for Visual Understanding. Built upon a visual encoder [33, 34] and an LLM [35, 36, 6], multimodal LLMs [1, 2, 3, 4, 37, 38, 39, 40] produce language responses based on visual inputs, allowing new capabilities like visual reasoning, dialogue, and instruction following. Despite their powerful visual understanding capabilities, most existing multimodal LLMs are limited to text outputs. Therefore, equipping these LLMs with visual generation ability would be the key step towards next-generation multimodal intelligence. In this work, we build OpenUni upon InternVL3 [37], adapting its original world knowledge to image generation.

Text-to-Image Generation with Diffusion Models. Diffusion models [14, 9, 15, 11, 13, 16, 41, 42] have become the dominant paradigm for image generation, producing high-quality visual content conditioned on language descriptions. Pioneering works [9, 11] typically formulate image generation as Denoising Diffusion Probabilistic Models [43] (DDPM), based on a U-Net [44] architecture. In more recent works [45, 46, 47, 48], Flow Matching (FM) forgoes explicit diffusion simulation by learning an ODE-driven continuous transformation from noise to data, effectively subsuming diffusion processes as a special case and permitting more direct probability transport paths to improve efficiency. For model architecture, state-of-the-art frameworks [41, 48, 49, 47, 46, 50] replace the U-Net backbone with diffusion transformers [17] (DiTs). Among these models, SANA [46, 50] enhances both training and sampling efficiency with increased compression ratio [51] and linear attention [52]. In this work, we choose SANA as the diffusion module of OpenUni to reduce computation cost.

Frozen LLMs. Freezing the pre-trained weights of LLMs and incorporating task-specific modules has been an economic and effective approach to expanding their functionality [53, 54, 31, 29]. As an early attempt, F-LMM [53] builds a mask head on top of frozen LLMs, endowing the LLMs with grounding ability while preserving their reasoning and instruction-following capabilities. For image generation, LlamaFusion [54] introduces extra transformer modules for visual generation, alongside LLMs' original architecture. More recently, MetaQuery [31] and BLIP3-o [29] effectively bridge multimodal LLMs and diffusion models with a set of learnable queries. Inspired by these works, OpenUni is built upon frozen multimodal LLMs.

Unifying Multimodal Understanding and Generation. There are two design philosophies regarding unifying multimodal understanding and generation. One subtrack [19, 20, 21, 22, 23, 24, 25] explores native multimodal models

Table 1: Architecture specifications and number of training images of OpenUni, MetaQuery [31] and BLIP3-o [29]. *In BLIP3-o, the DiT that predicts CLIP features is regarded as the connector.

Model	MLLM	#Connector Params	Diffusion Model	#Images
MetaQuery-B [31]	LLaVA-OV-0.8B [4]	316M	SANA-1.6B-512 [46]	25M
MetaQuery-L [31]	Qwen2.5VL-3B [56]	Unknown	SANA-1.6B-512 [46]	25M
MetaQuery-XL [31]	Qwen2.5VL-7B [56]	Unknown	SANA-1.6B-512 [46]	25M
BLIP3-o-4B [29]	Qwen2.5VL-3B [56]	1.4B*	SDXL (2.6B) [11]	30M
BLIP3-o-8B [29]	Qwen2.5VL-7B [56]	1.4B*	SDXL (2.6B) [11]	30M/50M
OpenUni-B-512	InternVL3-1B [37]	54M	SANA-0.6B-512 [46]	23M
OpenUni-L-512	InternVL3-2B [37]	225M	SANA-1.6B-512 [46]	23M
OpenUni-L-1024	InternVL3-2B [37]	225M	SANA-1.5-1.6B-1024 [50]	23M

Table 2: Detailed hyperparameters in pre-training (stage I) and fine-tuning (stage II).

Setting	Stage I	Stage II
Diffusion Model	Frozen	Trainable
Learning Rate	10^{-4}	10^{-5}
Batch Size	512	256
Optimizer	AdamW [58]	AdamW [58]
Grad. Clip	1.0	1.0
Weight Decay	0.05	0.05
Betas	(0.9, 0.95)	(0.9, 0.95)
Schedule	Cosine	Cosine
Training Steps	100,000	10,000
Warm-up Steps	1,000	100

from scratch and shares the parameters of LLM for both tasks. This type of work typically struggles to accommodate the two inherently heterogeneous tasks that require representations at different levels of granularity. Disentangled visual encoders [22, 23] or mixture of experts [24, 25, 55] are usually adopted to handle conflicting pathways.

The more resource-efficient approach stitches well-trained LLMs and generation models to build unified frameworks, connecting them with intermediate ViT features [26, 27, 28, 30] or learnable queries [31, 29]. Among these works, MetaQuery [31] and BLIP3-o [29] build unified frameworks upon frozen multimodal LLMs, effectively transferring the knowledge learned in understanding tasks to visual generation. Our OpenUni follows the simple architecture introduced by MetaQuery [31] and achieves significantly higher performance with fewer learnable parameters, setting a clean and strong baseline for this research direction.

3 OpenUni

3.1 Model

Architecturally, OpenUni follows the design of MetaQuery [31], comprising N learnable queries, a multimodal LLM, a transformer-based connector, and a diffusion model. In our implementation, we set $N = 256$. The visual understanding ability of the multimodal LLM is fully retained since its weights remain frozen. During image generation, the learnable queries extract conditioning information from the user's prompt during the LLM's forward pass; this information is then processed by the connector and passed to the diffusion model via its cross-attention module.

Lightweight Connector. The architecture of OpenUni's connector is adapted from SigLIP's visual encoder [34]. Different from pioneering works [31, 29] that feature a heavy connecting module between the LLM and diffusion model, OpenUni's connector only comprises six transformer layers.

Model Variants. In this work, we build three model variants. OpenUni-B-512 is based on InternVL3-1B [37] and SANA-0.6B-512px [46] while OpenUni-L-512 adopts InternVL3-2B [37] and SANA-0.6B-512px [46]. In addition, we increase the resolution of OpenUni-L's image generation by changing the diffusion model to SANA-1.5-1.6B-1024px [50]. This higher-resolution variant is named OpenUni-L-1024. The model specifications of OpenUni are provided in Table 1, alongside comparisons with the pioneering works MetaQuery [31] and BLIP3-o [29].

Prompt Format. For text-to-image generation, we use the following prompt template to format user instruction: "`User: Generate an image <caption>\n Assistant:`". `<caption>` represents the image description. During training, `<caption>` is randomly set to empty for 10% data samples to enable classifier-free guidance (CFG) [57] in inference.

Table 3: Results from the GenEval benchmark for text-to-image generation. Here, BLIP3-o-8B* indicates the model that is trained with 30 million additional proprietary data samples. We highlight the best results in **bold**.

Type	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	LlamaGen [63]	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [64]	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [64]	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt-α [65]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [64]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [15]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [66]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [11]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [16]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
	SD3-Medium [45]	**0.99**	**0.94**	0.72	0.89	0.33	0.60	0.74
Unified	Chameleon [19]	-	-	-	-	-	-	0.39
	SEED-X [67]	0.97	0.58	0.26	0.80	0.19	0.14	0.51
	LMFusion [54]	-	-	-	-	-	-	0.63
	Show-o [21]	0.95	0.52	0.49	0.82	0.11	0.28	0.68
	EMU3 [66]	-	-	-	-	-	-	0.66
	TokenFlow-XL [68]	0.95	0.60	0.41	0.81	0.16	0.24	0.63
	Janus [22]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Janus-Pro-1B [23]	0.98	0.82	0.51	0.89	0.65	0.56	0.73
	Janus-Pro-7B [23]	**0.99**	0.89	0.59	0.90	0.79	0.66	0.80
	Harmon-0.5B [30]	**0.99**	0.80	0.57	0.87	0.55	0.48	0.71
	Harmon-1.5B [30]	**0.99**	0.86	0.66	0.85	0.74	0.48	0.76
	MetaQuery-B [31]	-	-	-	-	-	-	0.74
	MetaQuery-L [31]	-	-	-	-	-	-	0.78
	MetaQuery-XL [31]	-	-	-	-	-	-	0.80
	BLIP3-o-4B [29]	-	-	-	-	-	-	0.81
	BLIP3-o-8B [29]	-	-	-	-	-	-	0.83
	BLIP3-o-8B* [29]	-	-	-	-	-	-	0.84
	OpenUni-B-512	**0.99**	0.91	0.74	0.90	0.77	0.73	0.84
	OpenUni-L-512	**0.99**	0.91	**0.77**	0.90	0.75	0.76	0.85
	OpenUni-L-1024	**0.99**	0.92	0.76	**0.91**	**0.82**	**0.77**	**0.86**

3.2 Training Recipe

We adopt a two-stage training strategy, where we first align the LLM and the diffusion model in a pre-training stage. Then we fine-tune the aligned modules using high-quality training data. The training hyperparameters are listed in Table 2.

Stage 1: Pre-training. The primary goal of this stage is to train the learnable queries and the lightweight connector to effectively bridge the multimodal LLM and the diffusion transformer. Parameters of both the LLM and the diffusion model are frozen in this stage. The connector learns to translate the LLM's output features (elicited by the 256 learnable queries) into conditioning signals that the diffusion model can interpret. We use a large composite dataset comprising several publicly available image/text collections: text-to-image-2M[59], LAION-Aesthetic- 6M[60], Megalith-10M[61], RedCaps-5M[62]. All of these images are captioned by LLMs. This results in a pre-training corpus of roughly 23 million image-text pairs.

Stage 2: High-Quality Finetuning. To refine the generative capabilities of the entire system (connector and diffusion model) for improved instruction adherence, image quality, and robustness to diverse prompts, we leverage the instruction tuning dataset released by BLIP3-o [29]. The dataset consists of 60,000 high-quality image-text pairs generated by prompting GPT-4o with diverse captions and using models like DALL-E3 and Midjourney for image synthesis.

4 Evaluation

This section details the evaluation setup, benchmarks, and results for OpenUni. We evaluated OpenUni's capabilities in both image generation and multimodal understanding, comparing them against state-of-the-art models. Our evaluation aims to demonstrate OpenUni's ability to achieve competitive performance with a simpler and light-weight architecture.

Table 4: Results from the DPG-Bench for text-to-image generation.Here, BLIP3-o-8B* indicates that the model is trained with 30 million additional proprietary data samples. We highlight the best results in **bold**.

Type	Method	Global	Entity	Attribute	Relation	Other	Overall↑
Gen. Only	SDv1.5 [64]	74.63	74.23	75.39	73.49	67.81	63.18
	PixArt-α [65]	74.97	79.32	78.60	82.57	76.96	71.11
	Lumina-Next [12]	82.82	88.65	86.44	80.53	81.82	74.63
	SDXL [11]	83.27	82.43	80.91	86.76	80.41	74.65
	Playground v2.5 [69]	83.06	82.59	81.20	84.08	83.50	75.47
	Hunyuan-DiT [42]	84.59	80.59	88.01	74.36	86.41	78.87
	PixArt-Σ [13]	86.89	82.89	88.94	86.59	87.68	80.54
	Emu3-Gen [66]	85.21	86.68	86.84	90.22	83.15	80.60
	DALL-E 3 [16]	**90.97**	89.61	88.39	**90.58**	89.83	83.50
	SD3-Medium [45]	87.90	**91.01**	88.83	80.70	88.68	84.08
Unified	Show-o [21]	-	-	-	-	-	67.27
	Janus [22]	82.33	87.38	87.70	85.46	86.41	79.68
	Janus-Pro-1B [23]	87.58	88.63	88.17	88.98	88.30	82.63
	Janus-Pro-7B [23]	86.90	88.90	89.40	89.32	89.48	**84.19**
	MetaQuery-B [31]	-	-	-	-	-	80.04
	MetaQuery-L [31]	-	-	-	-	-	81.10
	MetaQuery-XL [31]	-	-	-	-	-	82.05
	BLIP3-o-4B [29]	-	-	-	-	-	79.36
	BLIP3-o-8B [29]	-	-	-	-	-	80.73
	BLIP3-o-8B* [29]	-	-	-	-	-	81.60
	OpenUni-B-512	85.87	87.33	86.54	86.91	89.43	80.29
	OpenUni-L-512	81.37	87.67	88.64	88.18	**89.77**	81.54
	OpenUni-L-1024	87.01	90.02	**89.63**	90.28	88.62	83.08

Table 5: Results from the WISE benchmark evaluating world knowledge in text-to-image generation. Here, BLIP3-o-8B* indicates the model that is trained with an additional 30 million proprietary data. We highlight the best results in **bold**.

Type	Method	Cultural	Time	Space	Biology	Physics	Chemistry	Overall↑
Gen. Only	SDv1.5 [64]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
	SDv2.1 [64]	0.30	0.38	0.35	0.33	0.34	0.21	0.32
	Emu3-Gen [66]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
	FLUX.1-schnell [47]	0.39	0.44	0.50	0.31	0.44	0.26	0.40
	SD3-Medium [45]	0.42	0.44	0.48	0.39	0.47	0.29	0.42
	SDXL [11]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
	SD3.5-Large [45]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
	PixArt-α [65]	0.45	0.50	0.48	**0.49**	0.56	0.34	0.47
	FLUX.1-dev [47]	0.48	**0.58**	0.62	0.42	0.51	0.35	0.50
Unified	Show-o [21]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
	Janus [22]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
	Janus-Pro-1.5B [23]	0.20	0.28	0.45	0.24	0.32	0.16	0.26
	MetaQuery-B [31]	0.44	0.49	0.58	0.41	0.49	0.34	0.46
	MetaQuery-L [31]	**0.56**	0.57	0.62	0.48	**0.63**	0.42	0.55
	MetaQuery-XL [31]	**0.56**	0.55	0.62	0.49	**0.63**	**0.41**	0.55
	Harmon-1.5B [30]	0.38	0.48	0.52	0.37	0.44	0.29	0.41
	BLIP3-o-4B [29]	-	-	-	-	-	-	0.50
	BLIP3-o-8B [29]	-	-	-	-	-	-	0.52
	BLIP3-o-8B* [29]	-	-	-	-	-	-	**0.62**
	OpenUni-B-512	0.37	0.45	0.58	0.39	0.50	0.30	0.43
	OpenUni-L-512	0.51	0.49	0.64	0.48	**0.63**	0.35	0.52
	OpenUni-L-1024	0.49	0.53	**0.69**	**0.49**	0.56	0.39	0.52

Table 6: Results on image understanding benchmarks. Since the parameters of the MLLM (InternVL3 [37]) are frozen, OpenUni preserves its excellent performance on the following benchmarks. We highlight the best results in **bold**.

Model	MMBench	SEED	MM-Vet	MME-P	MMMU	RWQA	TEXTVQA	POPE
EMU2 Chat [26]	-	62.8	48.5	-	34.1	-	66.6	-
Chameleon-7B [19]	19.8	27.2	8.3	202.7	22.4	39.0	0.0	-
Chameleon-34B [19]	32.7	-	9.7	604.5	38.8	39.2	0.0	-
Seed-X [67]	70.1	66.5	43.0	1457.0	35.6	-	-	-
VILA-U [20]	-	59.0	33.5	1401.8	-	46.6	48.3	85.8
LMFusion [54]	72.1	63.7	-	1603.7	41.7	60.0	-	-
Show-o-512 [21]	-	-	-	1097.2	26.7	-	-	73.8
EMU3 [66]	58.5	68.2	37.2	-	31.6	57.4	64.7	85.2
MetaMorph [28]	75.2	71.8	-	-	-	58.3	60.5	-
TokenFlow-XL [73]	76.8	72.6	48.2	1551.1	43.2	56.6	77.6	86.8
Janus-1.3B [22]	69.4	63.7	34.3	1338.0	30.5	-	-	87.0
Janus-Pro-7B [23]	79.2	72.1	50.0	1567.1	41.0	-	-	-
Harmon-0.5B [30]	59.8	62.5	-	1148.0	34.2	-	-	86.5
Harmon-1.5B [30]	65.5	67.1	-	1155.0	38.9	-	-	87.6
MetaQuery-B [31]	58.5	66.6	29.1	1238.0	31.4	-	-	-
MetaQuery-L [31]	78.6	73.8	63.2	1574.3	53.1	-	-	-
MetaQuery-XL [31]	**83.5**	76.9	**66.6**	**1685.2**	**58.6**	-	-	-
BLIP3-O-4B [29]	78.6	73.8	60.1	1527.7	46.6	60.4	78.0	-
BLIP3-O-8B [29]	**83.5**	**77.5**	**66.6**	1682.6	50.6	**69.0**	**83.1**	-
OpenUni-B (InternVL3-1B [37])	72.6	58.2	59.5	1491.22	43.4	58.2	74.1	**90.7**
OpenUni-L (InternVL3-2B [37])	81.1	64.6	62.2	1626.88	48.6	64.3	77.0	89.6

4.1 Image Generation

To assess the text-to-image generation capabilities of OpenUni, we employ a range of established benchmarks focusing on prompt adherence, semantic alignment, and world knowledge. Specifically, **GenEval** [70] is employed to evaluate the model's proficiency in following complex textual prompts, focusing on generating images with correct object attributes, counts, positions, and colors. Results are reported across various categories such as single object, multiple objects, counting, colors, and position. We also use **DPG-Bench** [71], a benchmark designed to examine the intricate semantic alignment capabilities of text-to-image models using lengthy and dense prompts. For DPG-bench, we report scores across its defined categories (Global, Entity, Attribute, Relation, Other) and the overall score. Finally, **WISE** [72] is employed to evaluate the model's incorporated world knowledge and reasoning capability within the context of image generation.

The performance of OpenUni on the GenEval benchmark is presented in Table 3. It is remarkable that our smallest variant OpenUni-B-512 (0.84) already matches the performance of larger models like MetaQuery-XL (0.80) and BLIP3-o-8B (0.84). OpenUni-L-1024 archives an overall score of 0.86. Table 4 summarizes the performance of OpenUni on the DPG-Bench benchmark. On DPG-Bench, OpenUni-L-1024 obtains an overall score of 83.08, surpassing all model variants of MetaQuery and BLIP3-o while being comparable to Janus-Pro-7B. For world knowledge evaluation on the WISE benchmark (Table 5), OpenUni-L-512/1024 achieves 0.52, already matching the performance of BLIP3-o-8B trained on 30M public data. Finally, we visualize the image generation results in Figure 2.

4.2 Multimodal Understanding

Since the frozen InternVL3 [37] models are used to build OpenUni, their core understanding capabilities are primarily inherited. We summarize their performance on several standard multimodal understanding benchmarks and compare with mainstream unified models. The reported benchmarks include MMBench, SEED-Bench, MM-Vet, MME-Perception (MME-P), MMMU, RealWorldQA (RWQA) and TextVQA. Here, we choose MMBench for evaluating diverse tasks requiring perception and reasoning; SEED-Bench for assessing generative comprehension; MM-Vet for evaluating integrated capabilities of large multimodal models; MME-Perception (MME-P) as a comprehensive benchmark for perception capabilities; MMMU for massive multi-discipline multimodal understanding and reasoning; RealWorldQA (RWQA) for assessing performance on real-world question answering; TextVQA which focuses on visual question answering where answers are present as text in the image; and POPE for evaluating object hallucination. As shown in Table 6, OpenUni achieves competitive performance on these established benchmarks with only 1B and 2B activated parameters, thanks to InternVL3's outstanding visual perception and reasoning ability. Finally, we show some examples of OpenUni-L performing image understanding tasks in Figure 3.

Figure 2: OpenUni is able to generate high-quality images and precisely follow user instructions. The above results are obtained by OpenUni-L-1024.

5 Discussion & Limitations

In this work, we have introduced OpenUni, a simple but strong baseline for building unified models upon existing multimodal LLMs. As an ongoing project, OpenUni has the following limitations that will be addressed in future works: 1) the current OpenUni models struggle to render texts in generated images; 2) our largest model is based on a 2B LLM and 1.6B diffusion model. We believe scaling up the model sizes would further improve both understanding and generation performance of OpenUni; 3) image-to-image generation tasks (e.g., reconstruction and editing) are left for future updates.

User: What is the time now?

Model Output: The time on clock is approximately 10:10.

User: Describe the image in detail.

Model Output: In the im a young child is sitting on a wooden deck, reaching out to feed a capybara. The child is wearing a gray long-sleeve shirt and blue jeans. The capybara, a large rodent, is standing on the deck and eating food, which includes pieces of fruit. The child is holding an orange stick, possibly to offer the capybara more food. In the background, there are people standing, and part of a modern black structure is visible. The scene appears to be outdoors, likely in a park or zoo setting.

User: What is the third benefit?

Model Output: The third benefit is "Enhances Writing Capability."

User: What main text are there in the image?

Model Output: The main text in the image is "Serving Soul since + Twenty Twelve."

User: Where is this scene? Describe the scene in detail.

Model Output: This scene is from Singapore, specifically showcasing the iconic Marina Bay Sands and Gardens by the Bay. Here's a detailed description:

1. **Marina Bay Sands**: The three tall towers connected by a horizontal structure on top are part of Marina Bay Sands. This is a well-known landmark in Singapore, known for its unique design and the observation deck on the top.

2. **Gardens by the Bay**: To the left of Marina Bay Sands, you can see the distinctive, flower-shaped structures known as the Flower Dome and Cloud Forest. These are part of the Gardens by the Bay, a major horticultural attraction in Singapore.

3. **Waterfront and Bridges**: The area is surrounded by water, with a bridge visible in the foreground. This is part of the Marina Bay area, which is known for its extensive waterfront and transportation infrastructure.

4. **Skyline and Background**: The background shows a view of the sea with several ships, indicating the proximity to the shipping lanes and ports of Singapore. The sky has a warm, golden hue, suggesting it is either sunrise or sunset.

5. **Urban Layout**: The layout includes modern buildings, green spaces, and a mix of commercial and recreational areas, typical of Singapore's urban planning.

This scene captures the essence of Singapore's modern skyline and its blend of natural beauty and architectural marvels.

Figure 3: OpenUni has inherited InternVL3 [37]'s excellent multimodal understanding ability to recognize visual patterns and comprehend general world knowledge. The results are obtained by OpenUni-L (based on InternVL3-2B).

References

[1] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

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

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[6] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

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

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[10] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[11] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[12] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. *arXiv preprint arXiv:2406.18583*, 2024.

[13] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Sigma: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

[14] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[16] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[19] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[20] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.

[21] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[22] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

[23] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025.

[24] Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. *arXiv preprint arXiv:2412.09604*, 2024.

[25] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

[26] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[27] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024.

[28] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

[29] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025.

[30] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025.

[31] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025.

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[37] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

[38] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[39] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[40] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[42] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

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

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[45] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

[46] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024.

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

[48] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models wfith scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[49] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *arXiv preprint arXiv:2406.18583*, 2024.

[50] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.

[51] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024.

[52] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17302–17313, 2023.

[53] Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. *arXiv preprint arXiv:2406.05821*, 2024.

[54] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Llamafusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

[55] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *arXiv preprint arXiv:2504.07951*, 2025.

[56] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[57] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[58] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[59] Jacky He and contributors. text-to-image-2M: A high-quality, diverse text–image training dataset. `https://huggingface.co/datasets/jackyhate/text-to-image-2M`, 2024.

[60] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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.

[61] Ollin Matsubara and Draw Things AI Team. Megalith-10M: A dataset of 10 million public-domain photographs. `https://huggingface.co/datasets/madebyollin/megalith-10m`, 2024. CC0/Flickr-Commons images; Florence-2 captions available in the *megalith-10m-florence2* variant.

[62] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image–text data created by the people, for the people. In *NeurIPS, Datasets and Benchmarks Track*, 2021.

[63] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[65] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[66] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[67] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[68] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

[69] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

[70] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[71] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

[72] Yuwei Niu, Munan Ning, Mengren Zheng, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[73] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.