# Ovis-U1 Technical Report

**Ovis Team, Alibaba Group**

https://github.com/AIDC-AI/Ovis-U1
https://huggingface.co/AIDC-AI/Ovis-U1-3B

## Abstract

In this report, we introduce Ovis-U1, a 3-billion-parameter unified model that integrates multimodal understanding, text-to-image generation, and image editing capabilities. Building on the foundation of the Ovis series, Ovis-U1 incorporates a diffusion-based visual decoder paired with a bidirectional token refiner, enabling image generation tasks comparable to leading models like GPT-4o. Unlike some previous models that use a frozen MLLM for generation tasks, Ovis-U1 utilizes a new unified training approach starting from a language model. Compared to training solely on understanding or generation tasks, unified training yields better performance, demonstrating the enhancement achieved by integrating these two tasks. Ovis-U1 achieves a score of 69.6 on the OpenCompass Multi-modal Academic Benchmark, surpassing recent state-of-the-art models such as Ristretto-3B and SAIL-VL-1.5-2B. In text-to-image generation, it excels with scores of 83.72 and 0.89 on the DPG-Bench and GenEval benchmarks, respectively. For image editing, it achieves 4.00 and 6.42 on the ImgEdit-Bench and GEdit-Bench-EN, respectively. As the initial version of the Ovis unified model series, Ovis-U1 pushes the boundaries of multimodal understanding, generation, and editing.


Figure 1: Comprehensive illustration of the functional capabilities of Ovis-U1.

## 1 Introduction

The rapid evolution of multimodal large language models (MLLMs) has been a driving force behind the increasing sophistication of artificial general intelligence (AGI). Recent developments, notably GPT-4o, introduced by OpenAI (2025), have shown that unified models capable of both understanding and generating across multiple modalities can significantly transform a wide range of real-world applications. GPT-4o integrates native image generation with advanced language capabilities, empowering users to execute complex visual tasks through natural language dialogue. These tasks (e.g., image editing (Brooks et al., 2023), multi-view synthesis (Mildenhall et al., 2021), style transfer (Gatys et al., 2016), object detection (Zou et al., 2023), instance segmentation (Hafiz & Bhat, 2020), depth estimation (Mertan et al., 2022), normal estimation (Qi et al., 2018)), which previously required specialized models, can now be performed with high efficiency and accuracy. This represents a breakthrough in multimodal perception and marks the beginning of a new era where unified multimodal understanding and generation models (Zhang et al., 2025) handle both text and visual tasks seamlessly.

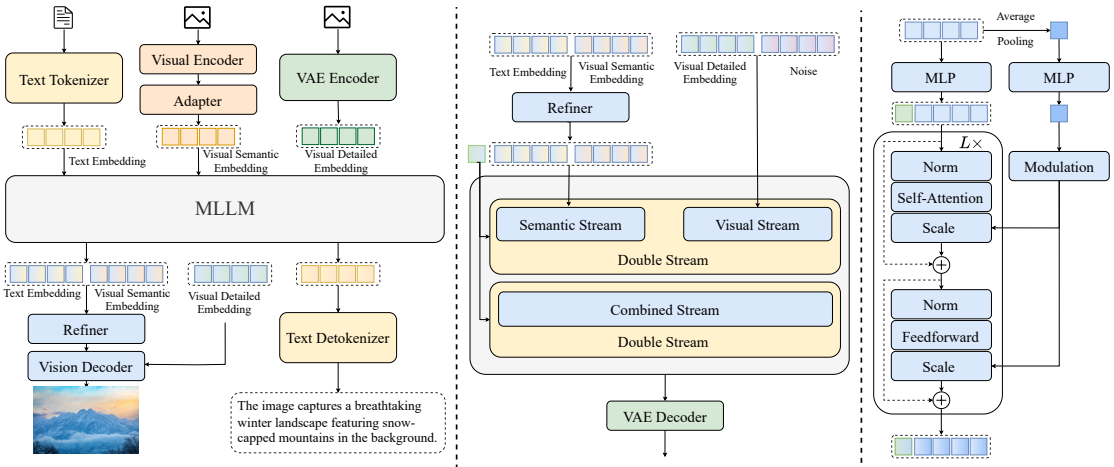
The emergence of GPT-4o marks a significant transition towards a unified multimodal understanding and generation framework in areas related to AGI. This raises two fundamental questions. First, how can a multimodal understanding model be endowed with the capability to generate images? This requires careful design of a visual decoder that can work seamlessly with the large multimodal language model. Second, how can a unified model be effectively trained on both understanding and generation tasks? We have observed that GPT-4o's understanding performance is enhanced by integrating image generation capabilities, suggesting that unified training may collaboratively improve performance across a range of tasks. In this report, we will study these two questions by our Ovis-U1 model.

Drawing inspiration from GPT-4o, we present Ovis-U1, a unified model with 3 billion parameters that expands the capabilities of the Ovis series (Lu et al., 2024b). This model incorporates a novel visual decoder built on a diffusion Transformer architecture (Labs, 2024a; Esser et al., 2024) and a bidirectional token refiner (Ma et al., 2024; Kong et al., 2024) to enhance the interaction between textual and visual embeddings. These advancements allow Ovis-U1 to generate high-quality images from textual descriptions and refine images based on textual prompts. Ovis-U1 is trained using a unified strategy that simultaneously tackles various tasks with a diverse array of multimodal data. Comprehensive ablation studies show that our unified training approach collaboratively enhances both understanding and generation performance.

The vision for Ovis-U1 is twofold: firstly, to advance existing MLLM models by introducing novel architecture and training strategies that improve the understanding, generation, and editing of multimodal data, thereby enhancing precision and flexibility in handling complex tasks. Secondly, by open-sourcing Ovis-U1, we aim to accelerate AI development within the community, encouraging collaborative research and innovation to hasten the creation of general-purpose AI systems capable of advanced multimodal reasoning and manipulation.

In this report, the emergence of Ovis-U1 represents a significant step forward in the development of multimodal AI systems, expanding on the strengths of the Ovis series while paving the way for future advancements. Below, we show the key features of Ovis-U1:

- **Diversity of Data**: Ovis-U1 has been trained on a diverse composition of multimodal data, spanning text-image understanding, text-to-image generation, and image editing tasks. This diverse training ensures that the model excels across a wide range of applications, from generating detailed images from textual descriptions to refining and editing images based on complex prompts. By learning from multiple tasks in a unified framework, Ovis-U1 achieves improved generalization, seamlessly handling real-world multimodal challenges with high accuracy.
- **Architecture Improvement**: Building upon the previous Ovis models, Ovis-U1 enhances its multimodal understanding capabilities by introducing a novel visual decoder based on diffusion architecture and a bidirectional token refiner to strengthen the interaction between textual and visual embeddings. The visual decoder utilizes multimodal diffusion Transformer (MMDiT) with rotary position embedding (RoPE) as the backbone, allowing for high-fidelity image generation from text. The bidirectional token refiner improves the interaction between textual and visual features, significantly enhancing text-to-image synthesis and image manipulation tasks.
- **Unified Training**: Unlike previous models that specialized in single tasks, Ovis-U1 adopts a unified training approach that leverages multimodal capabilities across 6 training stages, as show in Table 2. This approach ensures that the model learns to balance and integrate knowledge across various tasks—ranging from understanding textual and visual inputs to generating and editing images. This unified framework enables Ovis-U1 to perform seamlessly across different use cases, further pushing the boundaries of multimodal AI performance.




(a) Ovis-U1 Architecture (b) Visual Decoder Architecture (c) Refiner Architecture

Figure 2: The overall architecture of Ovis-U1. (a) The Ovis-U1 model integrates both textual and visual inputs through a shared Multimodal Large Language Model (MLLM), using a visual decoder for image generation and a text detokenizer for text generation. An adapter bridges the vision encoder with the MLLM. A refiner module enhances the quality of the conditional embedding before decoding. (b) The architecture of the refiner module consists of two stacked Transformer blocks with modulation applied to the average pooled features. The green token represents a learnable *[CLS]* token used to aggregate global information from the conditional embeddings.

Table 1: The model structure details of Ovis-U1.

| Module | #Param. (M) | Pretrain | Trained by Stage |
|---|---|---|---|
| LLM | 1720 | Qwen/Qwen3-1.7B | 3 |
| Vision Decoder | 1046 | - | 0, 4, 5 |
| Visual Encoder | 578 | apple/aimv2-large-patch14-448 | 2, 3 |
| Adapter | 135 | - | 1, 2, 3 |
| VAE | 84 | madebyollin/sdxl-vae-fp16-fix | |
| Refiner | 81 | - | 0, 4, 5 |
| Total | 3644 | | |

## 2 Architecture

The structure of Ovis-U1 is presented by Fig. 2. The details of each module are summarized in Tab. 1. Overall, Ovis-U1 inherits the architecture of Ovis (Lu et al., 2024b) by adding a visual decoder to generate the image.

**LLM & Text tokenizer.** We utilize the Qwen3 series (Yang et al., 2025) as the backbone for the large language model. To create a unified model with 3 billion parameters, we employ Qwen3-1.7B. Unlike previous approaches that directly use a multimodal large language model (such as Qwen-VL (Bai et al., 2025)) as the backbone and keep it unchanged during training, our Oivs-U1 is initialized with a language model and trained using both visual understanding and generation data. This unified training approach enhances the model's performance in both understanding and generation tasks collaboratively.

**Visual Encoder & Adapter.** We enhance the visual encoder from Ovis and adopt its original visual adapter. The visual encoder, initialized from Aimv2-large-patch14-448 (Fini et al., 2024), is modified to natively handle images of arbitrary resolutions, avoiding the sub-image partitioning strategy. To achieve this, we adapt the original fixed-size positional embeddings via interpolation and incorporate 2D Rotary Positional Embeddings (RoPE) (Su et al., 2024) for improved spatial awareness. The architecture also employs a variable-length sequence attention mechanism (Dao et al., 2022; Dao, 2024), following the token packing strategy from NaViT (Dehghani et al., 2023) to efficiently process batches of images with varying resolutions. Following the encoder, a visual adapter bridges the vision and language modalities using the identical probabilistic tokenization scheme from Ovis. This module uses a pixel shuffle operation for spatial compression, followed by a linear head and a softmax function to transform features into a probability distribution over a visual vocabulary. The final embedding, fed to the LLM, is a weighted

average from a learnable embedding table based on this distribution.

**Visual Decoder & VAE.** We use a diffusion transformer as the visual decoder. Specifically, inspired by FLUX Labs (2024a), we use MMDiT with RoPE (Su et al., 2024) as the backbone and flow matching as the training target. By decreasing the number of layers and attention heads from 57 and 24 to 27 and 16, respectively, a 1B visual decoder is obtained. This decoder is initialized randomly and trained from scratch. Due to the limited capacity of the decoder, we employ the VAE model from SDXL with 4 channels and freeze it during the unified training. In line with FLUX.1 Redux (Labs, 2024b), the visual semantic embedding is concatenated with the text embedding to serve as semantic conditions for image generation. Additionally, following FLUX.1 Kontext (Labs, 2025), the context image is encoded into latent tokens using the VAE encoder. Compared to the visual semantic embedding, these context image tokens contain detailed information from the context image. Finally, these visually detailed embeddings, along with the image tokens (Noise), are input into the decoder's visual stream.

**Refiner.** We introduce a bidirectional token refiner to promote the interaction between visual embedding and textual embedding. Following Kong et al. (2024); Ma et al. (2024), we stack 2 transformer blocks with the modulation mechanism to consist of our refiner. Since different layers of LLM capture different levels of information about images and texts, in order to make full use of the differences in information granularity at different layers, we propose to concatenate the features of the last layer with the features of the second-to-last layer and then send them to the refiner for information interaction, which helps to generate better conditional guidance. It is worth noting that the previous text-based generation model FLUX (Labs, 2024a) usually introduces CLIP to capture global features. In order to replace CLIP (Radford et al., 2021), we introduce learnable *[CLS]* token. By concatenating the learnable *[CLS]* token and the embedding generated by LLM, and then sending them to the refiner for interaction, global information can be captured.

## 3 Data Composition and Training Procedure

### 3.1 Data Composition

To train Ovis-U1, we leverage three distinct types of multimodal data: multimodal understanding data, text-to-image generation data, and image+text-to-image generation data. Below, we elaborate on each category.

**Multimodal Understanding Data**. This dataset consists of both publicly available and in-house developed data. The public datasets we utilize include COYO (Byeon et al., 2022), Wukong (Gu et al., 2022), Laion (Schuhmann et al., 2022), ShareGPT4V (Chen et al., 2024a), and CC3M (Sharma et al., 2018). Additionally, we have established a data preprocessing pipeline to filter out noisy data, enhances caption quality, and adjusts data ratio to ensure optimal training performance.

**Text-to-Image Generation Data**. For our text-to-image generation tasks, we draw from the Laion5B dataset (Schuhmann et al., 2022) and JourneyDB (Sun et al., 2023). Specifically, with Laion5B, we first select samples with an aesthetic score above 6. We then employ the Qwen model (Wang et al., 2024) to generate detailed descriptions for each selected image, culminating in the creation of the Laion-aes6 dataset.

**Image+Text-to-Image Generation Data**. This category can be further subdivided into four specific types:

- Image Editing Data: We utilize public datasets including OmniEdit (Wei et al., 2024), UltraEdit (Zhao et al., 2024), and SeedEdit (Ge et al., 2024).
- Reference-image-driven image generation Data: Our sources for this include Subjects200K (Tan et al., 2024) and SynCD (Kumari et al., 2025) for subject-driven image generation and StyleBooth (Han et al., 2024) for style-driven image generation.
- Pixel-Level Controlled Image Generation Data: This encompasses tasks such as canny-to-image, depth-to-image, inpainting, and outpainting, drawing from MultiGen_20M (Qin et al., 2023).
- In-House Data: We have also constructed additional datasets to complement publicly available resources, incorporating style-driven data, content removal, style translation, de-noise/de-blur data, colorization data, text rendering data, etc.

### 3.2 Training Procedure

Different from previous works that directly using the pretrained MLLM (e.g., Qwen-VL (Bai et al., 2025)), we train our model from the pretrained LLM. Given the pretrained LLM and visual encoder, Ovis has 4

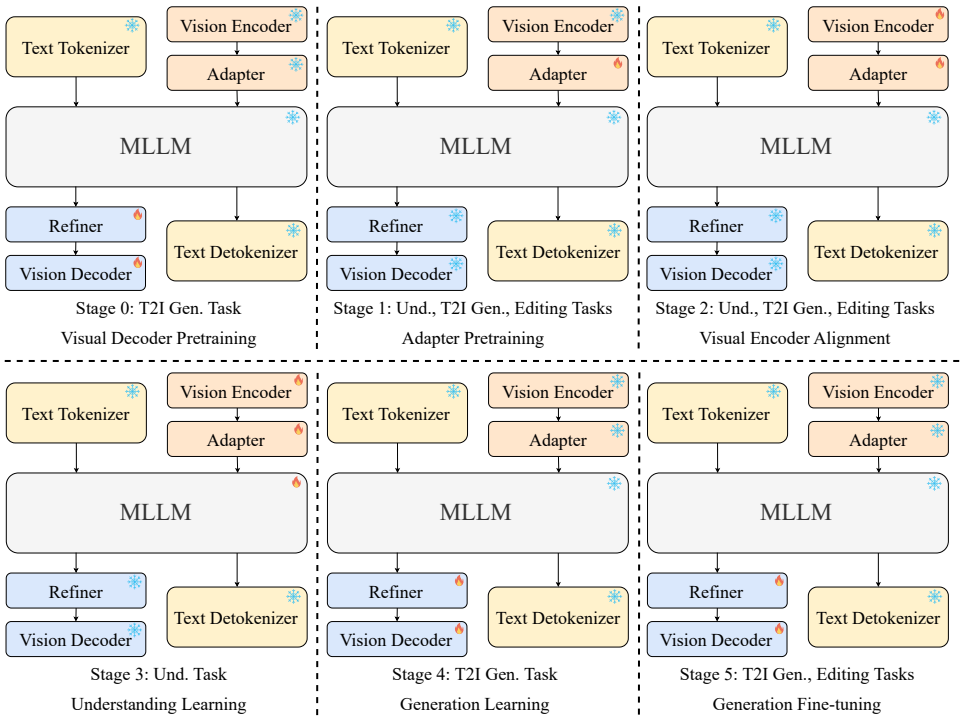



Figure 3: Overview of the proposed six-stage training pipeline. We progressively train the Ovis-U1 model through a sequence of carefully designed stages. Snowflake and flame icons denote frozen and trainable components, respectively.

Table 2: Details about each training stage of Ovis-U1.

| Stage | Trained Param. | Task | Steps (K) | Batch Size | Learning Rate |
|---|---|---|---|---|---|
| 0 | refiner + $Dec_i$ | T2I Gen. | 500 | 1024 | 1e-4 |
| 1 | adapter | Und., T2I Gen., Editing | 1.51 | 8192 | 5e-4 |
| 2 | $Enc_i$ + adapter | Und., T2I Gen., Editing | 2.63 | 8192 | 1e-4 |
| 3 | $Enc_i$ + adapter + LLM | Und. | 23 | 2240 | 5e-5 |
| 4 | refiner + $Dec_i$ | T2I Gen. | 275 | 256 | 5e-5 |
| 5 | refiner + $Dec_i$ | T2I Gen., Editing | 325 | 256 | 5e-5 |

training procedures totally: adapter pretraining, visual encoder alignment, understanding learning, and DPO. We add more training stages for generation. Details of each training stage are presented in Tab. 2.

**Stage 0: Visual Decoder Pretraining.** We construct a 1B diffusion transformer for the visual decoder, starting with random initialization and training it from scratch to develop basic image generation capabilities. This stage uses text-to-image training data, enabling the visual decoder, along with the refiner, to generate images from LLM embeddings.

**Stage 1: Adapter Pretraining.** The adapter serves as a bridge between the visual encoder and LLM, aligning visual and textual embeddings. More details are provided in the Ovis paper (Lu et al., 2024b). The adapter is randomly initialized and requires training during this stage. Unlike Ovis, Ovis-U1 is trained across understanding, text-to-image, and image editing tasks.

**Stage 2: Visual Encoder Alignment.** In this stage, both the visual encoder and adapter are fine-tuned together to further align visual and textual embeddings. Similar to Stage 1, this stage employs all three tasks for training, with the generation task aiding in the alignment of embeddings from different modalities.

**Stage 3: Understanding Learning.** This stage is the same as that of Ovis, where the parameters of the visual encoder, adapter, and LLM are trained on understanding tasks. Following this stage, these parameters are fixed to preserve the understanding capability.

**Stage 4: Generation Learning.** Since Stage 3 tunes the LLM parameters, we subsequently train the refiner and visual decoder to align with the optimized text and image embeddings. Our experiments indicate an improvement in text-to-image performance compared to Stage 0, as Stages 1-3 refine text embeddings to

Table 3: Performance of unified models on understanding, text-to-image generation and image editing. † refers to rewriting the prompt. × indicates the model is incapable of performing the task. ‡ indicates results from our own tests.

| **Model** | #Params. | Understanding | | | Text-to-image Generation | | Image Editing | |
|---|---|---|---|---|---|---|---|---|
| | | MMB | MMMU | MMVet | GenEval | DPG-Bench | ImgEdit | GEdit-EN |
| GPT-4o | - | 86.0 | 72.9 | 76.9 | 0.84 | - | 4.2 | 7.53 |
| Janus-Pro | 7B | 75.5 | 36.3 | 39.8 | 0.80 | 84.19 | × | × |
| Emu3 | 8B | 58.5 | 31.6 | 37.2 | 0.54 | 80.60 | × | × |
| BLIP3-o 4B | 3B + 1.4B | 78.6 | 46.6 | 60.1 | 0.81† | 79.36 | × | × |
| BLIP3-o 8B | 7B + 1.4B | 83.5 | 58.6 | 66.6 | 0.84† | 81.60 | × | × |
| BAGEL | 7B + 7B | 85.0 | 55.3 | 67.2 | 0.82 | 85.07 | 3.20 | 6.52 |
| UniWorld-V1 | 7B + 12B | 83.5 | 58.6 | 67.1 | 0.84† | 81.38 | 3.26 | 4.85 |
| OmniGen | 3.8B | × | × | × | 0.68 | 81.16 | 2.96 | 5.06 |
| OmniGen2 | 3B + 4B | 79.1 | 53.1 | 61.8 | 0.86† | 83.57 | 3.44 | 6.42 |
| OmniGen2‡ | 3B + 4B | 76.8 | 51.2 | 58.5 | - | - | - | - |
| Ovis-U1 | 2.4B + 1.2B | 77.8 | 51.1 | 66.7 | 0.89 | 83.72 | 4.00 | 6.42 |

better align with image embeddings.

**Stage 5: Generation Fine-tuning.** Building on text-to-image capabilities, the final training stage involves fine-tuning the decoder for text-to-image and image editing tasks.

## 4 Evaluation

Like GPT-4o, recent unified multimodal models possess the ability to comprehend input images, generate images based on input prompts, and edit images according to instructions. Therefore, we benchmark the models on three tasks: image understanding, text-to-image generation, and image editing.

**Understanding.** To evaluate the understanding capabilities, we use the widely-used OpenCompass Multi-modal Academic Benchmarks[1], including MMBench (Liu et al., 2024a), MMStar (Chen et al., 2024b), MMMU-Val (Yue et al., 2024), MathVista-Mini (Lu et al., 2024a), HallusionAvg (Guan et al., 2024), AI2D-Test (Kembhavi et al., 2016), OCRBench (Liu et al., 2024b), MMVet (Yu et al., 2024). The Avg Score is obtained by averaging the performance over these 8 benchmarks. Most powerful multimodal large language models have been evaluated on this Benchmark. Therefore, the unified model can compare with them conveniently.

**Text-to-Image Generation.** To evaluate the text-to-image generatation capability, we use CLIPScore (Hessel et al., 2021), DPG-Bench (Hu et al., 2024), and GenEval (Ghosh et al., 2023) benchmarks. CLIPScore was used in DALL-E 3 (Betker et al., 2023) and the first 1K prompts[2] are intended to be used for CLIPScore calculation. DPG-Bench and GenEval are two widely-used benchmarks for text-to-image models and unified models. Some previous works rewrite prompts of GenEval to boost the performance. In this paper, we report the performance with the raw prompts.

**Image Editing.** To evaluate the image editing capability, we employ GEdit-Bench (Liu et al., 2025) and ImgEdit (Ye et al., 2025), two recently introduced benchmarks featuring 606 and 811 image-instruction pairs, respectively. Both benchmarks utilize the advanced GPT model to evaluate the edited images.

## 5 Experiments

In this section, we begin by summarizing the overall performance of Ovis-U1 across understanding tasks, text-to-image generation, and image editing capabilities. Following this, we present several ablation studies to demonstrate the effectiveness of our proposed methodologies, particularly focusing on the refiner design and the performance improvements achieved through collaborative training of understanding and generation components. Finally, we showcase qualitative results to illustrate our model's capabilities.

[1] https://rank.opencompass.org.cn/leaderboard-multimodal/

[2] https://github.com/openai/dalle3-eval-samples/blob/main/prompts/8k_coco.txt

Table 4: Evaluation of understanding ability on OpenCompass Multi-modal Academic Benchmarks.

| **Model** | MMB | MMS | MMMU | MathVista | Hallusion | AI2D | OCRBench | MMVet | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | 86 | 70.2 | 72.9 | 71.6 | 57 | 86.3 | 82.2 | 76.9 | 75.4 |
| InternVL2.5-2B | 70.9 | 54.3 | 43.2 | 51.1 | 42.3 | 74.9 | 80.2 | 62.6 | 59.9 |
| SAIL-VL-2B | 73.7 | 56.5 | 44.1 | 62.8 | 45.9 | 77.4 | 83.1 | 44.2 | 61 |
| InternVL3-2B | 78 | 61.1 | 48.7 | 57.6 | 41.9 | 78.6 | 83.1 | 67 | 61.1 |
| Qwen2.5-VL-3B | 76.8 | 56.3 | 51.2 | 61.2 | 46.6 | 81.4 | 82.8 | 60 | 64.5 |
| Ovis2-2B | 76.9 | 56.7 | 45.6 | 64.1 | 50.2 | 82.7 | 87.3 | 58.3 | 65.2 |
| SAIL-VL-1.5-2B | 78.5 | 62.6 | 46.4 | 67 | 50 | 83.7 | 89.1 | 58.8 | 67 |
| Ristretto-3B | 80.2 | 62.8 | 51.3 | 67.6 | 50.2 | 84.2 | 84.7 | 60.7 | <u>67.7</u> |
| Ovis-U1 | 77.8 | 61.3 | 51.1 | 69.4 | 56.3 | 85.6 | 88.3 | 66.7 | **69.6** |

Table 5: Evaluation of text-to-image generation ability on GenEval. † denotes using the rewritten prompts

| **Model** | Single object | Two object | Counting | Colors | Position | Attribute binding | Overall |
|---|---|---|---|---|---|---|---|
| GPT-4o | 0.99 | 0.92 | 0.85 | 0.92 | 0.75 | 0.61 | 0.84 |
| BAGEL | 0.99 | 0.94 | 0.81 | 0.88 | 0.64 | 0.63 | 0.82 |
| BAGEL† | 0.98 | 0.95 | 0.84 | 0.95 | 0.78 | 0.77 | 0.88 |
| UniWorld-V1 | 0.99 | 0.93 | 0.79 | 0.89 | 0.49 | 0.70 | 0.80 |
| UniWorld-V1† | 0.98 | 0.93 | 0.81 | 0.89 | 0.74 | 0.71 | 0.84 |
| OmniGen | 0.98 | 0.84 | 0.66 | 0.74 | 0.40 | 0.43 | 0.68 |
| OmniGen2 | 1 | 0.95 | 0.64 | 0.88 | 0.55 | 0.76 | 0.80 |
| OmniGen2† | 0.99 | 0.96 | 0.74 | 0.98 | 0.71 | 0.75 | <u>0.86</u> |
| Ovis-U1 | 0.98 | 0.98 | 0.90 | 0.92 | 0.79 | 0.75 | **0.89** |

Table 6: Evaluation of text-to-image generation ability on DPG-Bench.

| **Model** | Global | Entity | Attribute | Relation | Other | Overall |
|---|---|---|---|---|---|---|
| BAGEL | 88.94 | 90.37 | 91.29 | 90.82 | 88.67 | **85.07** |
| UniWorld-V1 | 83.64 | 88.39 | 88.44 | 89.27 | 87.22 | 81.38 |
| OmniGen | 87.90 | 88.97 | 88.47 | 87.95 | 83.56 | 81.16 |
| OmniGen2 | 88.81 | 88.83 | 90.18 | 89.37 | 90.27 | 83.57 |
| Ovis-U1 | 82.37 | 90.08 | 88.68 | 93.35 | 85.20 | <u>83.72</u> |

Table 7: Evaluation of image editing ability on ImgEdit-Bench.

| **Model** | Add | Adjust | Extract | Replace | Remove | Background | Style | Hybrid | Action | **Overall** |
|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | 4.61 | 4.33 | 2.9 | 4.35 | 3.66 | 4.57 | 4.93 | 3.96 | 4.89 | 4.2 |
| MagicBrush | 2.84 | 1.58 | 1.51 | 1.97 | 1.58 | 1.75 | 2.38 | 1.62 | 1.22 | 1.90 |
| Instruct-P2P | 2.45 | 1.83 | 1.44 | 2.01 | 1.50 | 1.44 | 3.55 | 1.2 | 1.46 | 1.88 |
| AnyEdit | 3.18 | 2.95 | 1.88 | 2.47 | 2.23 | 2.24 | 2.85 | 1.56 | 2.65 | 2.45 |
| UltraEdit | 3.44 | 2.81 | 2.13 | 2.96 | 1.45 | 2.83 | 3.76 | 1.91 | 2.98 | 2.7 |
| OmniGen | 3.47 | 3.04 | 1.71 | 2.94 | 2.43 | 3.21 | 4.19 | 2.24 | 3.38 | 2.96 |
| Step1X-Edit | 3.88 | 3.14 | 1.76 | 3.40 | 2.41 | 3.16 | 4.63 | 2.64 | 2.52 | 3.06 |
| ICEdit | 3.58 | 3.39 | 1.73 | 3.15 | 2.93 | 3.08 | 3.84 | 2.04 | 3.68 | 3.05 |
| BAGEL | 3.56 | 3.31 | 1.7 | 3.3 | 2.62 | 3.24 | 4.49 | 2.38 | 4.17 | 3.2 |
| UniWorld-V1 | 3.82 | 3.64 | 2.27 | 3.47 | 3.24 | 2.99 | 4.21 | 2.96 | 2.74 | 3.26 |
| OmniGen2 | 3.57 | 3.06 | 1.77 | 3.74 | 3.2 | 3.57 | 4.81 | 2.52 | 4.68 | <u>3.44</u> |
| Ovis-U1 | 4.13 | 3.62 | 2.98 | 4.45 | 4.06 | 4.22 | 4.69 | 3.45 | 4.61 | **4.00** |

Table 8: Evaluation of image editing ability on GEdit-Bench-EN.

| Model | Background Change | Color Alteration | Material Modification | Motion Change | Portrait Beautification | Style Transfer | Subject Addition | Subject Removal | Subject Replacement | Text Modification | Tone Transformation | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | 7.205 | 6.491 | 6.607 | 8.096 | 7.768 | 6.961 | 7.622 | 8.331 | 8.067 | 7.427 | 8.301 | 7.534 |
| AnyEdit | 4.663 | 4.260 | 2.537 | 2.024 | 3.479 | 2.032 | 3.995 | 3.089 | 3.180 | 0.922 | 5.151 | 3.212 |
| Instruct-Pix2Pix | 3.825 | 5.182 | 3.688 | 3.509 | 4.339 | 4.560 | 3.461 | 2.031 | 4.237 | 0.955 | 4.733 | 3.684 |
| MagicBrush | 5.637 | 5.136 | 5.078 | 4.513 | 4.487 | 4.439 | 5.252 | 3.704 | 4.941 | 1.384 | 5.130 | 4.518 |
| OmniGen | 5.281 | 6.003 | 5.308 | 2.916 | 3.087 | 4.903 | 6.628 | 6.352 | 5.616 | 4.519 | 5.064 | 5.062 |
| Gemini | 6.781 | 6.369 | 6.040 | 6.938 | 5.591 | 4.676 | 7.501 | 6.447 | 7.003 | 5.765 | 6.350 | 6.315 |
| Step1X-Edit | 6.547 | 6.545 | 6.204 | 6.483 | 6.787 | 7.221 | 6.975 | 6.512 | 7.068 | 6.921 | 6.448 | 6.701 |
| Doubao | 7.430 | 7.095 | 6.339 | 6.973 | 6.972 | 6.767 | 7.674 | 6.748 | 7.447 | 3.471 | 7.383 | 6.754 |
| BAGEL | 7.324 | 6.909 | 6.381 | 4.753 | 4.573 | 6.150 | 7.896 | 7.164 | 7.021 | 7.320 | 6.218 | 6.519 |
| Ovis-U1 | 7.486 | 6.879 | 6.208 | 4.790 | 5.981 | 6.463 | 7.491 | 7.254 | 7.266 | 4.482 | 6.314 | **6.420** |

### 5.1 The main result

Table 3 summarizes the performance across multimodal understanding, generation, and editing tasks, comparing Ovis-U1 with models such as GPT-4o (OpenAI, 2025), Janus-Pro (Chen et al., 2025b), Emu3 (Wu et al., 2025b), BLIP3-o (Chen et al., 2025a), BAGEL (Deng et al., 2025), UniWorld-V1 (Lin et al., 2025), OmniGen (Xiao et al., 2024), and OmniGen2 (Wu et al., 2025a). Most of the results are sourced from OmniGen2, and we've independently tested OmniGen2's understanding capabilities. To ensure a fair comparison, we maintain consistent generation configurations across all benchmarks. Despite having only 3.34 billion parameters, Ovis-U1 shows outstanding performance across all tasks evaluated. Detailed results for each benchmark are presented in Tables 4 to 8.

Table 4 outlines the results on OpenCompass Multi-modal Academic Benchmarks. Beyond methodological differences, model size significantly impacts understanding capabilities. We compare Ovis-U1 with the leading models within the 3B parameter range, including InternVL2.5-2B (Chen et al., 2024c), SAIL-VL-2B (Dong et al., 2025), InternVL3-2B (Zhu et al., 2025), Qwen2.5-VL-3B (Bai et al., 2025), Ovis2-2B[3], SAIL-VL-1.5-2B[4], and Ristretto-3B[5]. Ovis-U1 surpasses all these models, setting a new benchmark for state-of-the-art performance, despite having only about 2B parameters dedicated to understanding tasks.

Tables 5 and 6 display the results for text-to-image generation capabilities on GenEval and DPG-Bench, respectively. We compare Ovis-U1 with recent open-source models, such as BAGEL (Deng et al., 2025), UniWorld-V1 (Lin et al., 2025), and OmniGen2 (Wu et al., 2025a). Ovis-U1 significantly outperforms OmniGen, despite having a similar number of parameters. It's noteworthy that Ovis-U1 is equipped with a 1B visual decoder, yet it achieves performance comparable to larger models.

Tables 7 and 8 present the results for image editing capabilities on ImgEdit-Bench and GEdit-Bench-EN, respectively. The performance of previous models on ImgEdit-Bench is referenced from OmniGen2 (Wu et al., 2025a), while the results for GEdit-Bench-EN are sourced from the leaderboard[6]. Our model demonstrates strong performance on both benchmarks.

### 5.2 Ablation study on refiner

As shown in Table 9, we explore various token refiner designs for text-to-image generation tasks, comparing both clip-based and clip-free approaches. It's important to note that these ablation studies were performed on earlier versions of our model and with a limited amount of training data. The baseline model, which combines the T5 text encoder (Raffel et al., 2020) with the CLIP image encoder trained on about 10M text-to-image data, demonstrated solid performance with a CLIPScore of 32.19 and a DPG-Bench score of 82.32. When T5 was replaced with Qwen2.5-1.5B-Instruct (Yang et al., 2024) in Variant V1, using only the last layer's features resulted in a performance degradation, with a CLIPScore of 32.12 and a DPG-Bench score of 80.97. However, concatenating the second-to-last and last-layer features in Variant V2 restored performance to baseline levels, with scores of 32.19 and 81.48, respectively. A further enhancement was achieved by replacing Qwen2.5-1.5B-Instruct with a version fine-tuned for image-text alignment (Ovis2) in Variant V3, which led to a slight improvement in DPG-Bench (82.37) but a minor drop in CLIPScore (32.18). Moreover, the clip-free approaches were tested, with Variant V5 using a CLS token for global information outperforming Variant V4, which used averaged refiner outputs. Despite the improvements, the clip-free variants still showed slightly lower performance compared to the baseline, suggesting the potential benefits of larger datasets for better exploration of clip-free methods.

[3] https://huggingface.co/AIDC-AI/Ovis2-2B

[4] https://huggingface.co/BytedanceDouyinContent/SAIL-VL-1d5-2B

[5] https://huggingface.co/LiAutoAD/Ristretto-3B

[6] https://step1x-edit.github.io/

Table 9: Effect of token refiner design. Each last row refers to our final solution.

| **Variant** | Data | CLIPScore | DPG-Bench | GenEval |
|---|---|---|---|---|
| Baseline (T5+CLIP) | ∼10M | 32.19 | 82.32 | 0.63 |
| (V1) Last-Layer Features | ∼10M | 32.12 | 80.97 | 0.62 |
| (V2) Concatenated Layer Features | ∼10M | 32.19 | 81.48 | 0.63 |
| (V3) Image-Text Aligned Features | ∼10M | 32.18 | 82.37 | 0.61 |
| (V4) Clip-Free - Averaged Features | ∼10M | 31.99 | 81.64 | 0.63 |
| (V5) Clip-Free - CLS Token | ∼10M | 32.13 | 81.91 | 0.61 |
| Baseline (T5+CLIP) | ∼50M | 32.57 | 82.97 | 0.69 |
| (V6) Clip-Free - Averaged Features | ∼50M | 32.47 | 82.65 | 0.71 |
| (V7) Clip-Free - CLS Token | ∼50M | 32.42 | 83.81 | 0.67 |

Table 10: Evaluation of understanding ability. Our unified training can enhance the understanding performance.

| **Variant** | MMB | MMS | MMMU | MathVista | Hallusion | AI2D | OCRBench | MMVet | **Avg** |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 75.8 | 55.4 | 45.0 | 63.0 | 47.9 | 82.3 | 87.4 | 49.8 | 63.33 |
| Unified Training | 77.0 | 56.9 | 44.6 | 66.3 | 46.7 | 83.7 | 87.9 | 52.8 | 64.47 |

When tested on the larger 50M training data, the baseline model again outperformed other designs, with a CLIPScore of 32.57 and a DPG-Bench score of 82.97. Among the clip-free designs, Variant V7 (using CLS tokens) achieved a higher DPG-Bench score of 83.81, though its CLIPScore was slightly lower than the baseline. These findings underscore the critical role of token refiner design in LLM-based text-to-image models, highlighting how careful selection of features, particularly in the token refinement process, significantly affects the alignment of text and image information, and consequently, model performance. The results suggest that further optimizations and larger datasets are necessary to fully realize the potential of clip-free methods, especially for improving generation performance on complex benchmarks.

### 5.3 Enhance understanding with unified training

Tab. 10 presents the detailed results from the OpenCompass Multi-modal Academic Benchmark. We use Ovis without the unified training as our baseline for comparison. Compared to this baseline, Ovis-U1 demonstrates a 1.14-point improvement in average score. This enhancement validates the effectiveness of leveraging text-to-image generation and image editing tasks for aligning the visual encoder during training stages 1 and 2. It is worth noting that most previous unified models typically underperform compared to their MLLM backbones. For instance, Ming-Lite-Uni (Gong et al., 2025), which utilizes Qwen2.5-VL-7B (Bai et al., 2025) as its backbone, achieves lower understanding performance. Some previous approaches (Chen et al., 2025a; Lin et al., 2025) keep the MLLM fixed, which misses the chance to enhance understanding performance.

### 5.4 Enhance generation with unified training

Tab. 11 and 12 summaries the image generation performance at various training stages. It's important to note that these ablation studies were performed on earlier versions of our model. In Stage 1, the model is trained using comprehensive text-to-image data, resulting in impressive performance outcomes. Following the alignment of visual and textual embeddings, Stage 4 further enhances the model's generation capabilities. In Stage 5, both text-to-image and image editing data are utilized. Notably, the inclusion of image editing data leads to a 0.77 improvement in the text-to-image performance on the DPG-Bench.

### 5.5 Instruction based image editing

Following the approach of InstructPix2Pix (Brooks et al., 2023), we implement classifier-free guidance for both text and image conditions. Fig. 4 illustrates the results of image editing under various classifier-free guidance settings. A higher $CFG_{img}$ value preserves more details from the input image in the generated output (see blue and green boxes in Fig. 4 for details), while a higher $CFG_{txt}$ value enhances the model's adherence to the editing instructions. Quantitative evaluation results are summarized in Table 13 and 14. Overall, our model demonstrates robustness to variations in the value of CFG, with result differences remaining within 0.2 for both ImgEdit-Bench and GEdit-Bench-EN. Additionally, the optimal CFG settings vary across different benchmarks. It's worth noting that our model achieves a score of 4.13 on

Table 11: Performance on DPG-Bench with different training stages. The generation ability is improved progressively across different training stages. Note that the results were obtained using an early version of our model.

| **Model** | Global | Entity | Attribute | Relation | Other | **Overall** |
|---|---|---|---|---|---|---|
| Stage1 | 80.85 | 89.56 | 89.22 | 93.54 | 80.00 | 83.81 |
| Stage4 | 84.50 | 90.40 | 89.92 | 94.12 | 78.40 | 84.66 |
| Stage5 | 83.59 | 90.81 | 89.46 | 94.31 | 80.80 | 85.43 |

Table 12: Performance on GenEval with different training stages. The generation ability is improved progressively across different training stages. Note that the results were obtained using an early version of our model.

| **Model** | Single object | Two object | Counting | Colors | Position | Attribute binding | **Overall** |
|---|---|---|---|---|---|---|---|
| Stage1 | 0.99 | 0.68 | 0.46 | 0.83 | 0.58 | 0.51 | 0.67 |
| Stage4 | 0.99 | 0.77 | 0.48 | 0.86 | 0.55 | 0.58 | 0.70 |
| Stage5 | 0.98 | 0.75 | 0.54 | 0.85 | 0.44 | 0.60 | 0.69 |

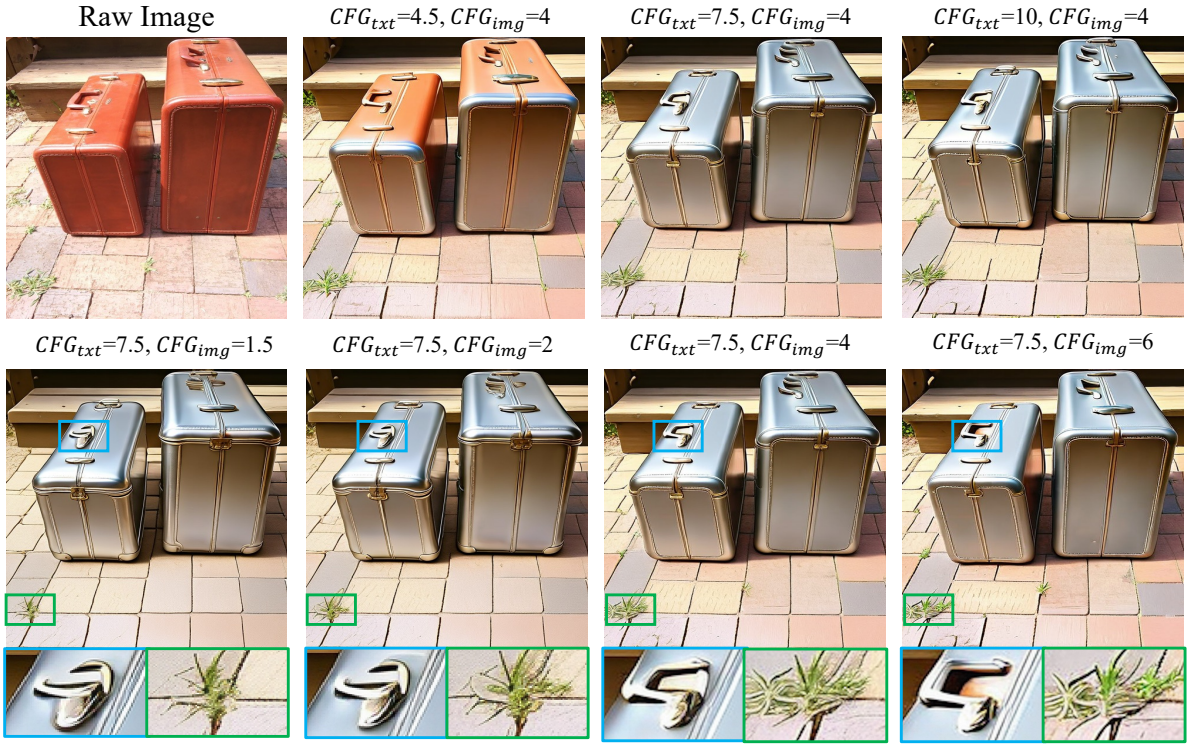



Figure 4: Qualitative results of different classifier free guidance on image editing. The instruction of editing is *"change the color of bags to silver"*.

Table 13: Performance on ImgEdit-Bench with different classifier free guidance. Note that the results were obtained using an early version of our model.

| $CFG_{img}$ | $CFG_{txt}$ | Add | Adjust | Extract | Replace | Remove | Background | Style | Hybrid | Action | **Overall** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.5 | 7.5 | 4.13 | 3.57 | 3.18 | 4.49 | 4.14 | 4.30 | 4.71 | 3.72 | 4.61 | 4.06 |
| 2 | 7.5 | 4.19 | 3.76 | 3.32 | 4.41 | 4.26 | 4.36 | 4.67 | 3.84 | 4.63 | 4.13 |
| 4 | 7.5 | 4.19 | 3.66 | 3.34 | 4.40 | 4.13 | 4.25 | 4.73 | 3.55 | 4.64 | 4.09 |
| 6 | 7.5 | 4.11 | 3.67 | 3.07 | 4.29 | 3.86 | 4.24 | 4.74 | 3.30 | 4.52 | 3.98 |
| 4 | 5 | 4.19 | 3.68 | 3.32 | 4.31 | 3.84 | 4.28 | 4.78 | 3.45 | 4.40 | 4.04 |
| 4 | 7.5 | 4.19 | 3.66 | 3.34 | 4.40 | 4.13 | 4.25 | 4.73 | 3.55 | 4.64 | 4.09 |
| 4 | 10 | 4.02 | 3.66 | 3.31 | 4.27 | 4.13 | 4.24 | 4.74 | 3.77 | 4.54 | 4.05 |

Table 14: Performance on GEdit-Bench-EN with different classifier free guidance. Note that the results were obtained using an early version of our model.

| $CFG_{img}$ | $CFG_{txt}$ | Background Change | Color Alteration | Material Modification | Motion Change | Portrait Beautification | Style Transfer | Subject Addition | Subject Removal | Subject Replacement | Text Modification | Tone Transformation | **Avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.5 | 7.5 | 7.538 | 6.265 | 6.095 | 4.607 | 5.909 | 6.660 | 6.702 | 6.902 | 7.029 | 4.120 | 5.761 | 6.144 |
| 2 | 7.5 | 7.674 | 6.544 | 6.064 | 4.536 | 5.696 | 6.394 | 6.668 | 6.945 | 7.203 | 4.320 | 6.208 | 6.205 |
| 4 | 7.5 | 7.706 | 7.056 | 6.177 | 4.248 | 5.770 | 6.595 | 7.143 | 7.209 | 7.002 | 4.449 | 6.499 | 6.351 |
| 6 | 7.5 | 7.573 | 6.829 | 6.025 | 3.915 | 5.636 | 6.508 | 7.166 | 6.857 | 7.344 | 4.327 | 6.620 | 6.254 |
| 4 | 5 | 7.768 | 7.020 | 6.477 | 3.874 | 5.772 | 6.697 | 7.116 | 7.236 | 7.105 | 4.254 | 6.823 | 6.377 |
| 4 | 7.5 | 7.706 | 7.056 | 6.177 | 4.248 | 5.770 | 6.595 | 7.143 | 7.209 | 7.002 | 4.449 | 6.499 | 6.351 |
| 4 | 10 | 7.654 | 6.762 | 6.074 | 4.654 | 6.083 | 6.535 | 7.252 | 7.248 | 7.185 | 4.222 | 6.216 | 6.353 |

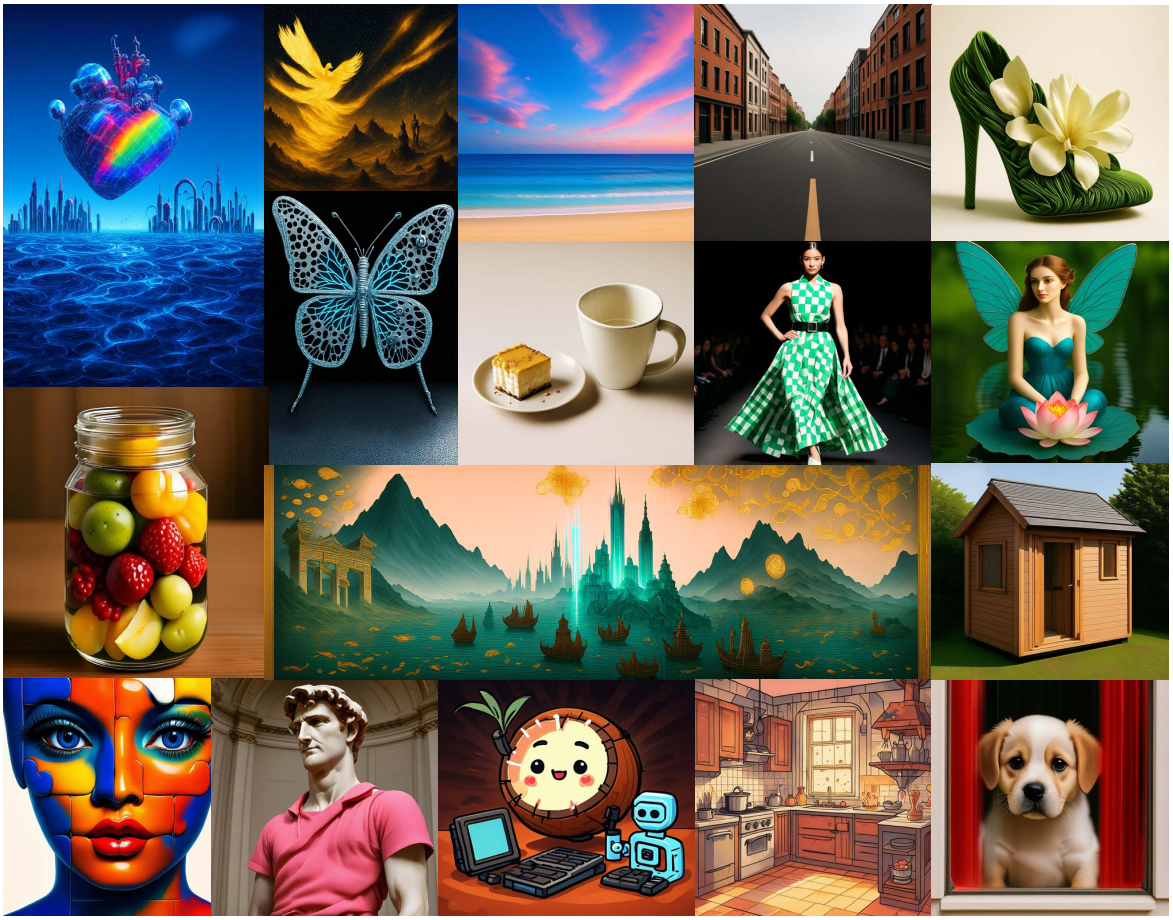


Figure 5: More qualitative results from Ovis-U1 on text-to-image generation.

ImgEdit-Bench with $CFG_{img}$ set to 2 and $CFG_{txt}$ set to 7.5, which is higher than the score reported in Tab. 3. This discrepancy is because, in Tab. 3, the same CFG settings are applied across all benchmarks.

### 5.6 More qualitative results

We present comprehensive qualitative results in Fig. 5, 6 and 7, to demonstrate Ovis-U1's capabilities across multimodal tasks.

As shown in Fig. 7, the model exhibits robust reasoning by interpreting complex visual content and generating contextually coherent responses. It excels at identifying fine-grained details, such as object attributes, spatial hierarchies, and subtle interactions while maintaining contextual awareness for domain-specific tasks.

Fig. 5 highlights the model's ability to produce high-fidelity images with diverse aesthetics and structural coherence. Ovis-U1 synthesizes visually striking outputs spanning photorealistic scenes, abstract concepts, and hybrid designs, while preserving intricate textures. It performs well even under complex prompts involving multi-object arrangements, spatial constraints, or abstract attribute bindings.


Figure 6: More qualitative results from Ovis-U1 on image editing.


Figure 7: More qualitative results from Ovis-U1 on image understanding.

Fig. 6 demonstrates Ovis-U1's precision in localized modifications while retaining background integrity. The model executes content replacement, stylistic transformations, and structural edits with minimal artifacts, adhering strictly to instructional prompts.

These qualitative results, paired with quantitative evaluations presented previously, position Ovis-U1 as a versatile foundation for multimodal generative tasks. Its compact 3.6 B parameter architecture balances efficiency and scalability, offering strong potential for performance gains through larger-scale training while maintaining practical deployment viability.

## 6 Conclusion

In this report, we present Ovis-U1, a 3-billion-parameter unified model that excels in multimodal understanding, text-to-image generation, and image editing. As the initial version in the Ovis unified model series, this report addresses key foundational challenges: the design of the visual decoder, its connector with large language models, and the comprehensive training procedure for the unified model. We emphasize the critical role of unified training in aligning the visual encoder, which significantly enhances both understanding and generation performance through collaborative training. Moreover, we utilize a robust evaluation framework for assessing the unified model's capabilities. We have curated widely-accepted benchmarks in the fields of understanding, text-to-image generation, and image editing to ensure comprehensive evaluation. With only 3B parameters, Ovis-U1 demonstrates strong performance across these benchmarks, even surpassing some task-specific models. This achievement underscores Ovis-U1's ability to advance the boundaries of unified model capabilities.

In the future, we will focus on advancing the powerful unified models. First, we plan to expand our model by increasing the number of parameters. In the realm of image generation, smaller models often struggle with artifacts and hallucinations. By incorporating more parameters, the model can mitigate these issues and produce higher quality images. Second, we will enhance our training data pipeline by collecting and curating more diverse, high-quality datasets specifically designed for unified model training, with particular emphasis on interleaved image-text content. Third, we plan to innovate architectural designs tailored for unified models. To enhance image editing capabilities, we will implement specialized visual encoder-decoder structures optimized to preserve fine-grained details from input images. Last but not least, we acknowledge that Ovis-U1 currently lacks a reinforcement learning stage, which has proven crucial for large model optimization. Developing effective methods to align unified multimodal models with human preferences remains an important open research question in this domain.

## 7 Contributors

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, Yang Li, Qing-Guo Chen

## References


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.

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.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

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. *arXiv preprint arXiv:2501.17811*, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders, 2024.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GENEVAL: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023.

Biao Gong, Cheng Zou, Dandan Zheng, Hu Yu, Jingdong Chen, Jianxin Sun, Junbo Zhao, Jun Zhou, Kaixiang Ji, Lixiang Ru, et al. Ming-Lite-Uni: Advancements in unified architecture for natural multimodal interaction. *arXiv preprint arXiv:2505.02471*, 2025.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020.

Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP (1)*, 2021.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. EllA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv:2403.05135*, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Nupur Kumari, Xi Yin, Jun-Yan Zhu, Ishan Misra, and Samaneh Azadi. Generating multi-image synthetic data for text-to-image customization. *arXiv preprint arXiv:2502.01720*, 2025.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024a.

Black Forest Labs. Introducing FLUX.1 tools. https://bfl.ai/announcements/24-11-21-tools, 2024b.

Black Forest Labs. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. https://bfl.ai/announcements/flux-1-kontext, 2025.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. UniWorld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024a.

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

Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Alican Mertan, Damien Jade Duff, and Gozde Unal. Single image depth estimation: An overview. *Digital Signal Processing*, 123:103441, 2022.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025.

Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 283–291, 2018.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678, 2023.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*, 2024.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025a.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025b.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, pp. 57730–57754. PMLR, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, 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.

Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.