Tuna: Taming Unified Visual Representations for Native Unified Multimodal Models

Zhiheng Liu[1,2,†], **Weiming Ren**[1,3,†], **Haozhe Liu**[1,4,‡], **Zijian Zhou**[1,‡], **Shoufa Chen**[1,‡], **Haonan Qiu**[1], **Xiaoke Huang**[1], **Zhaochong An**[1], **Fanny Yang**[1], **Aditya Patel**[1], **Viktar Atliha**[1], **Tony Ng**[1], **Xiao Han**[1], **Chuyan Zhu**[1], **Chenyang Zhang**[1], **Ding Liu**[1], **Juan-Manuel Perez-Rua**[1], **Sen He**[1], **Jürgen Schmidhuber**[4], **Wenhu Chen**[3], **Ping Luo**[2], **Wei Liu**[1], **Tao Xiang**[1], **Jonas Schult**[1,*], **Yuren Cong**[1,*]

[1]Meta BizAI, [2]HKU, [3]University of Waterloo, [4]KAUST
[†]Joint first authors, listed alphabetically by last name, [‡]Core contributors, [*]Joint project lead

Unified multimodal models (UMMs) aim to jointly perform multimodal understanding and generation within a single framework. We present Tuna, a native UMM that builds a unified continuous visual representation by cascading a VAE encoder with a representation encoder. This unified representation space allows end-to-end processing of images and videos for both understanding and generation tasks. Compared to prior UMMs with decoupled representations, Tuna's unified visual space avoids representation format mismatches introduced by separate encoders, outperforming decoupled alternatives in both understanding and generation. Moreover, we observe that stronger pretrained representation encoders consistently yield better performance across all multimodal tasks, highlighting the importance of the representation encoder. Finally, in this unified setting, jointly training on both understanding and generation data allows the two tasks to benefit from each other rather than interfere. Our extensive experiments on multimodal understanding and generation benchmarks show that Tuna achieves state-of-the-art results in image and video understanding, image and video generation, and image editing, demonstrating the effectiveness and scalability of its unified representation design.

Correspondence: zhiheng10528@connect.hku.hk, w2ren@uwaterloo.ca, schult@meta.com, yuren@meta.com
Project page: https://tuna-ai.org

Meta

1 Introduction

A long-term aspiration of multimodal AI is *natively*[1] unified multimodal generation, in which a single model can seamlessly understand and generate diverse modalities such as text, images, and videos. Recent advances in unified multimodal models (Team, 2024; Deng et al., 2025a; Xie et al., 2025a) have shown promising results towards this vision, suggesting that truly integrated multimodal intelligence is increasingly within reach.

A central challenge in developing native UMMs lies in how visual inputs are encoded into representations. Current UMMs adopt one of two approaches: (1) decoupled visual representations for understanding and generation tasks, or (2) a unified visual representation shared across both tasks. Intuitively, learning a unified visual representation for both tasks offers compelling advantages for UMMs.

First, UMMs with decoupled representations, such as BAGEL (Deng et al., 2025a) and Mogao (Liao et al., 2025), often adopt MoE-style architectures to handle different visual encoders, introducing additional parameters that increase training and inference costs. In contrast, a unified representation allows the model to operate within a single representation space, simplifying training and improving efficiency. Second, different vision encoders typically produce representations with incompatible formats. For the same input, features from a representation encoder (*e.g.* SigLIP (Zhai et al., 2023)) and a causal VAE encoder (*e.g.* Wan 2.1 VAE (Wan et al., 2025)) differ in (1) spatial compression (16× vs. 8×), (2) temporal compression (none vs. 4×), and (3) channel dimension (1152 vs. 16). These discrepancies can cause representation conflicts in decoupled models, whereas unified representations inherently avoid such inconsistencies. Finally, unified visual representations provide a clear pathway for achieving mutual enhancement between understanding and generation. While

[1]We define *native* UMMs as models that are pretrained jointly on both understanding and generation objectives, instead of assembling understanding-only and generation-only models with learnable connectors. We refer to the latter as *composite* UMMs.

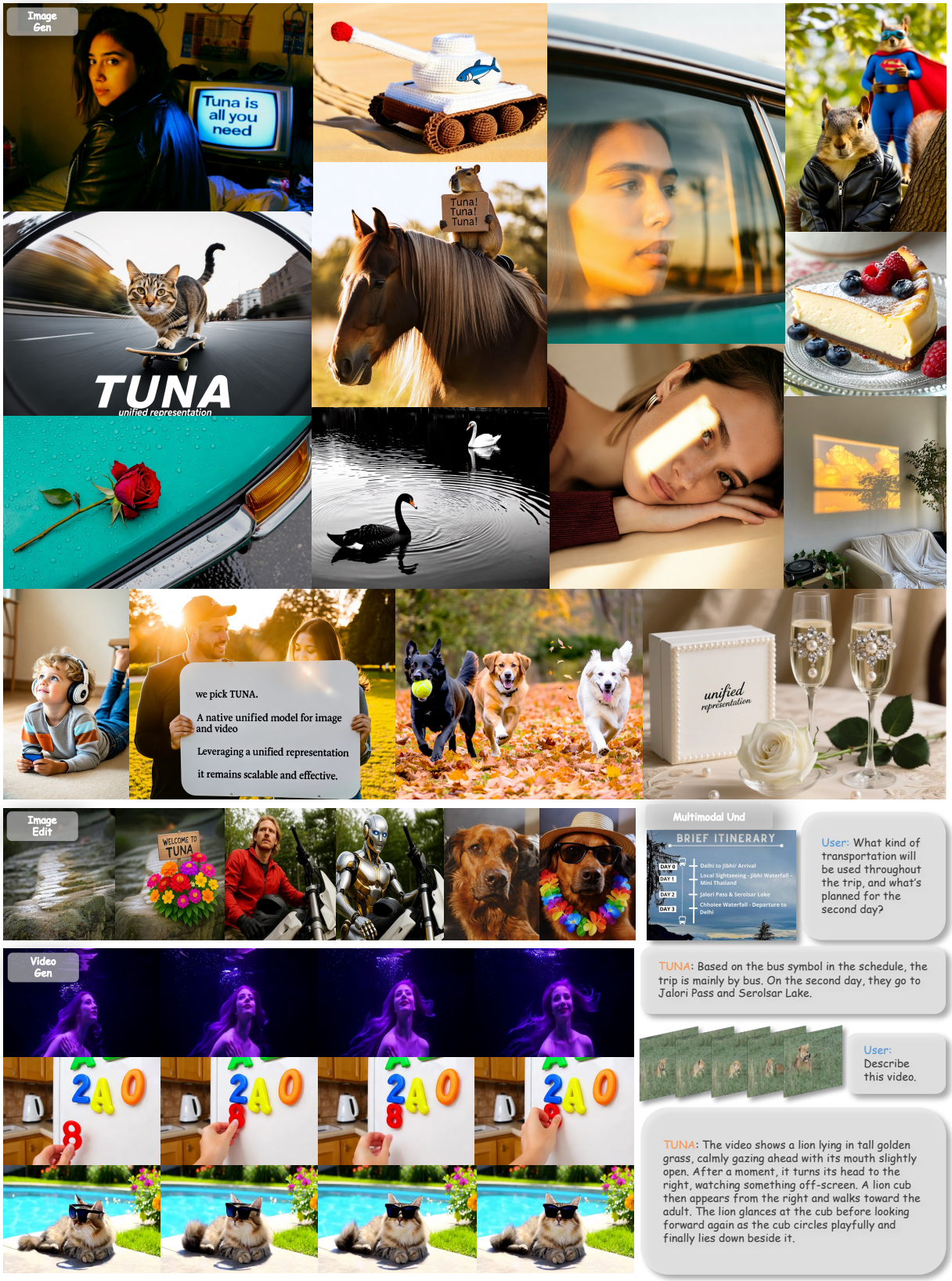


Figure 1 We present TUNA, a native unified multimodal model built on a unified visual representation, enabling diverse multimodal understanding and generation capabilities such as image and video understanding, image and video generation, and image editing.

recent studies like Ross (Wang et al., 2024a) and REPA (Yu et al., 2024) show task-specific improvements in understanding-only and generation-only models, this synergy remains underexplored in existing UMMs.

Nevertheless, current UMMs with unified visual representations often underperform their decoupled counterparts. Most existing approaches adopt a single type of vision encoder for both understanding and generation. For example, Chameleon (Team, 2024) and Transfusion (Zhou et al., 2024) use VQ-VAE (Esser et al., 2021), while Harmon (Wu et al., 2025d) utilizes the MAR encoder (Li et al., 2024c). This unified design tends to favor one task at the expense of the other. Show-o2 (Xie et al., 2025a) attempts to mitigate this issue by fusing SigLIP (Zhai et al., 2023) and VAE (Wan et al., 2025) features through a late-fusion strategy. However, our analysis in Section 3.4 reveals that its learned representation remains biased toward semantic features, resulting in limited generation quality.

To systematically address these limitations, we propose Tuna, a native UMM that employs unified visual representations across understanding and generation. Our design is simple yet highly effective: by directly connecting a VAE encoder to a representation encoder, we obtain representations that are sufficiently expressive for diverse multimodal tasks. These unified visual features are fused with text tokens and processed by an LLM decoder, which subsequently generates new text tokens and denoised images through autoregressive next-token prediction and flow matching. As illustrated in Figure 1, our unified visual representation enables Tuna to handle image and video understanding, image and video generation, and image editing within a single framework. By conducting a three-stage training, our model achieves state-of-the-art performance on multimodal understanding and generation benchmarks (*e.g.* 61.2% on MMStar (Chen et al., 2024b) and 0.90 on GenEval (Ghosh et al., 2023)).

Our main contributions can be summarized as follows:

1. We propose Tuna, a native unified multimodal model with unified visual representations, enabling image/video understanding, image/video generation, and image editing within a single framework.
2. Our extensive experiments show that Tuna's unified visual representation is highly effective, achieving state-of-the-art performance across multiple multimodal understanding and generation tasks.
3. We further perform a comprehensive ablation study, demonstrating the superiority of our unified visual representation design over existing methods such as Show-o2 and other models employing decoupled representations.

2 Our Method: Tuna

In this section, we introduce Tuna, a native unified multimodal model that employs unified visual representations across all multimodal understanding and generation tasks. We begin by outlining the key motivations behind our model design in Section 2.1, followed by a detailed description of Tuna's architecture and training pipeline in Sections 2.2 and 2.3, respectively. An overview of our overall framework is shown in Figure 2.

2.1 Motivation and Design Principles

We discuss the following observations, which motivate the design of Tuna and its unified visual representations:

Autoregressive vs. diffusion. Both text and image/video generation can be achieved using autoregressive (Grattafiori et al., 2024; Yang et al., 2025; Sun et al., 2024) or diffusion models (Nie et al., 2025; Batifol et al., 2025; Wan et al., 2025). In practice, leading understanding-only models (Bai et al., 2025; Wang et al., 2025c) adopt autoregressive models for text generation. On the other hand, state-of-the-art image and video generators (Esser et al., 2024; Wan et al., 2025) employ (latent[2]) diffusion models with flow matching.

Continuous vs. discrete visual representations. We observe that image and video generation models operating in a continuous (*e.g.*, KL-regularized) VAE latent space (Esser et al., 2024; Wan et al., 2025) outperform those using discrete representations (Sun et al., 2024), as discretization causes information loss and reduces fidelity. Similarly, multimodal understanding models (Bai et al., 2025; Wang et al., 2025c) typically rely on

[2]Due to higher generation fidelity and lower training costs, latent diffusion models are generally favoured over pixel diffusion (Zheng et al., 2025).

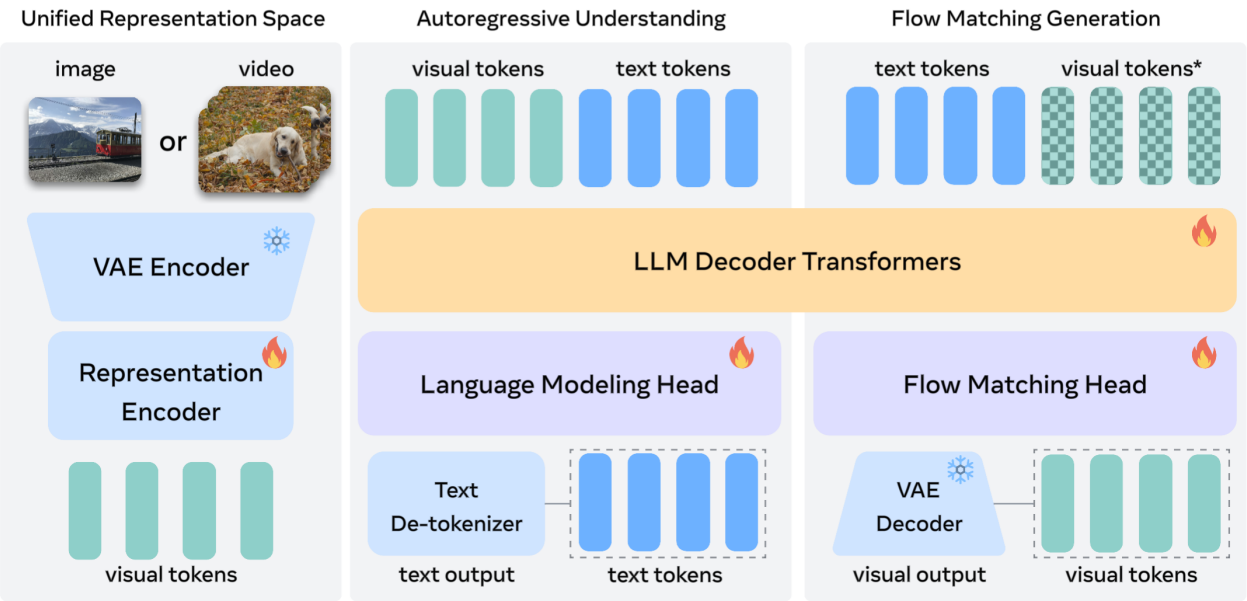


Figure 2 Overview of the TUNA architecture. Our model employs a VAE encoder and a representation encoder to construct unified visual representations, which are then combined with text tokens and processed by an LLM decoder. The decoder performs autoregressive text generation for understanding tasks and flow-matching-based visual generation for generation tasks. *During visual generation, noise is added to the visual tokens to enable diffusion-based generation.

continuous semantic features (*e.g.*, CLIP (Radford et al., 2021) features), suggesting that continuous visual representations are inherently more effective for both understanding and generation tasks.

Semantic representations benefit visual generation. Recent studies suggest that semantic features enhance visual generation. For instance, REPA (Yu et al., 2024) demonstrates that diffusion transformers benefit from aligning intermediate features with pretrained representation encoders like DINOv2 (Oquab et al., 2023). Concurrent to our work, RAE (Zheng et al., 2025) employs a frozen representation encoder to encode images into latent representations, showing that pretrained semantic features alone can reconstruct input images effectively.

VAE latents can also support understanding tasks. We observe that both discrete and continuous VAE latents, originally designed for visual reconstruction, can also support semantic understanding tasks. Recent approaches such as UniTok (Ma et al., 2025a) and TokLIP (Lin et al., 2025b) enhance VQ-VAE latents with semantic understanding capability through contrastive learning. Other works explore diffusion models with continuous VAE latents for semantic understanding and dense prediction tasks, including semantic segmentation (Zhu et al., 2024), object recognition (Li et al., 2023a), and image retrieval (Zuo et al., 2024).

Building on these observations, we design TUNA with the following key characteristics:

- TUNA integrates **autoregressive** text generation with **flow matching** for image and video generation.
- TUNA builds its unified visual representation on continuous VAE latents, as these latents effectively support both understanding and generation tasks.
- To further enhance performance, TUNA employs **representation encoders** to extract higher-level features from the VAE latents, improving the quality of both understanding and generation.

2.2 Model Architecture

Unified visual representations. As illustrated in Figure 2, TUNA constructs its unified visual representation using a VAE encoder and a representation encoder. Given an input image or video $\mathbf{X}$, we apply the 3D causal VAE encoder from Wan 2.2 (team, 2025), which downsamples the input by 16× spatially and 4× temporally, producing the latent $\mathbf{x}_1$. We then generate a noisy latent $\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$, where $t \in [0, 1]$ is a sampled timestep and $\mathbf{x}_0 \sim \mathcal{N}(0, 1)$. Next, we use the SigLIP 2 vision encoder Φ (patch size 16, pretrained resolution

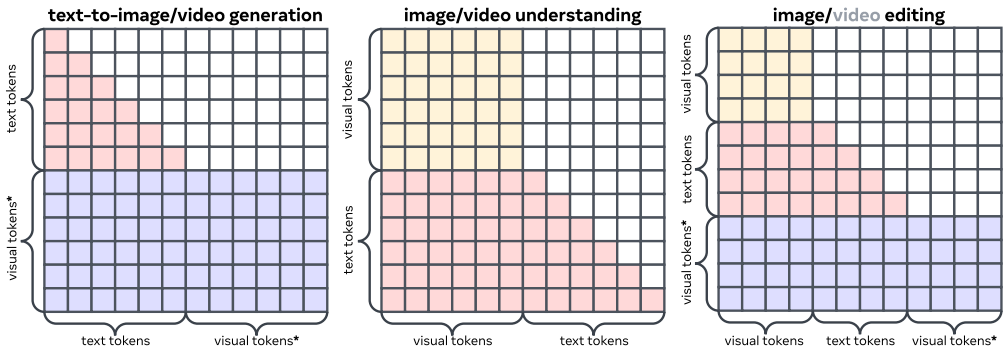


Figure 3 Attention masks in the LLM decoder for understanding and generation tasks. * indicates that the visual tokens are noised.

512) to extract semantic features from the VAE latents. Since the VAE encoder has 16× downsampling, we replace SigLIP 2's original 16×16 patch embedding layer with a randomly initialized 1×1 patch embedding layer, forming a modified encoder Φ'. This ensures that the token sequence lengths of $\Phi(\mathbf{X})$ and $\Phi'(\mathbf{x}_t)$ are consistent. Finally, we apply a two-layer MLP connector to obtain the unified visual representations $\mathbf{z} = \texttt{MLP}(\Phi'(\mathbf{x}_t))$. During training, we randomly sample t between $[0, 1]$ for visual generation and fix $t = 1$ for multimodal understanding such that $\mathbf{x}_t$ always corresponds to the clean latent.

For video inputs, where $\mathbf{x}_t \in \mathbb{R}^{b\times c\times f\times h\times w}$ (with b as batch size, f as the number of latent frames, and c, h, w as channel, height, and width), we aim to prevent the representation encoder Φ' from processing excessively long sequences. Instead of flattening all latent frames into a single sequence, we apply a window-based attention mechanism by reshaping the frame dimension into the batch dimension in Φ'. In `einops` notation, the unified visual representation $\mathbf{z}_v$ can be expressed as:

$$\bar{\mathbf{x}}_t = \texttt{rearrange}(\mathbf{x}_t, \texttt{b c f h w} \to (\texttt{b f})\ \texttt{c h w}), \tag{1}$$

$$\bar{\mathbf{z}}_v = \texttt{MLP}(\Phi'(\bar{\mathbf{x}}_t)) \in \mathbb{R}^{(b\times f)\times d}, \tag{2}$$

$$\mathbf{z}_v = \texttt{rearrange}(\bar{\mathbf{z}}_v, (\texttt{b f})\ \texttt{d} \to \texttt{b}\ (\texttt{f d})), \tag{3}$$

where d is the hidden dimension of the video tokens. This operation effectively allows Φ' to operate independently on each 4-frame window, significantly improving efficiency when processing video tokens.

LLM decoder and flow matching head. After obtaining the unified visual representation $\mathbf{z}$, we prepend a timestep token representing the sampled timestep t to $\mathbf{z}$, concatenate this visual token sequence with language tokens and feed the combined sequence into an LLM decoder (Qwen-2.5 (Bai et al., 2025)) for joint multimodal processing. Following standard UMM practices (Xie et al., 2024b; Deng et al., 2025a), we apply a causal attention mask on language tokens and a bidirectional attention mask on visual tokens within the LLM decoder layers, as illustrated in Figure 3. For multimodal understanding tasks, the LLM decoder output is passed through a language modeling head to generate text token predictions. For visual generation and image editing, we feed the full token sequence to a randomly initialized flow matching head to predict the velocity for flow matching. This head shares the LLM decoder architecture and adds timestep conditioning via AdaLN-Zero, following Show-o2 (Xie et al., 2025a) and DiT (Peebles and Xie, 2023). For generation and editing tasks, we adopt multimodal 3D-RoPE (Seawead et al., 2025; Su et al., 2024) over the concatenated text-visual sequence to handle interleaved instructions and visual content.

2.3 Training Pipeline

To effectively train our unified model, we adopt a three-stage training strategy that progressively adapts each model component to both understanding and generation tasks.

Stage 1: unified representation and flow matching head pretraining. In the first training stage, our objective is to adapt the semantic representation encoder to generate unified visual representations and to establish a robust initialization for the flow matching head. To this end, we train the representation encoder and flow matching head while freezing the LLM decoder, using two objectives: image captioning and text-to-image generation.

Models	Size	MME	GQA	RealWorldQA	SEED	MMMU	MMStar	AI2D	ChartQA	OCRBench
		perception	test-dev	test	image	val	avg	test	test	test
Understanding-only Models (LMMs)										
LLaVA-1.5 (Liu et al., 2023a)	7B	1510.7	62.0	54.8	65.8	35.7	33.1	55.5	17.8	31.8
Qwen-VL-Chat (Bai et al., 2023)	7B	1487.6	57.5	49.3	64.8	37.0	34.5	57.7	49.8	48.8
LLaVA-OV (Li et al., 2024a)	7B	1580.0	-	69.9	76.7	48.8	57.5	81.4	80.9	62.2
Composite UMMs										
TokenFlow-XL (Qu et al., 2025)	14B	1551.1	62.5	56.6	72.6	43.2	-	-	-	-
BLIP3-o (Chen et al., 2025a)	4B	1527.7	-	60.4	73.8	46.6	-	-	-	-
Tar (Han et al., 2025)	7B	1571.0	61.3	-	73.0	39.0	-	-	-	-
X-Omni (Geng et al., 2025)	7B	-	62.8	62.6	74.3	47.2	-	76.8	81.5	70.4
1.5B-scale Native UMMs										
Show-o (Xie et al., 2024b)	1.3B	1097.2	58.0	-	51.5	27.4	-	-	-	-
Harmon (Wu et al., 2025d)	1.5B	1155.0	58.9	49.8*	67.1	38.9	35.3*	57.0*	29.8*	11.2*
JanusFlow (Ma et al., 2025c)	1.3B	1333.1	60.3	41.2*	**70.5**	29.3	40.6*	54.2	42.4*	53.2*
SynerGen-VL (Li et al., 2025b)	2.4B	1381.0	-	-	-	34.2	-	-	-	-
Janus-Pro (Chen et al., 2025b)	1.5B	1444.0	59.3	52.6*	68.3	36.3	43.1*	64.5*	23.4	48.7
Show-o2 (Xie et al., 2025a)	1.5B	1450.9	60.0	56.5*	65.6	37.1	43.4	69.0	40.0*	24.5*
TUNA	1.5B	**1461.5**	**61.4**	**62.5**	69.3	**39.1**	**54.6**	**71.4**	**82.1**	**71.9**
7B-scale Native UMMs										
BAGEL (Deng et al., 2025a)	14B	1687.0	-	72.8	78.5	55.3	-	89.2	78.5	73.3
Emu3 (Wang et al., 2024c)	8B	-	60.3	57.4	68.2	31.6	-	70.0	-	68.7
VILA-U (Wu et al., 2024b)	7B	1401.8	60.8	-	59.0	-	-	-	-	-
MUSE-VL (Xie et al., 2024c)	7B	-	-	-	69.1	39.7	49.6	69.8	-	-
Janus-Pro (Chen et al., 2025b)	7B	1567.1	62.0	58.0*	72.1	41.0	48.3*	71.3*	25.8	59.0
Mogao (Liao et al., 2025)	7B	1592.0	60.9	-	74.6	44.2	-	-	-	-
Show-o2 (Xie et al., 2025a)	7B	1620.5	63.1	64.7*	69.8	48.9	56.6	78.6	52.3*	32.4*
TUNA	7B	**1641.5**	**63.9**	**66.1**	**74.7**	**49.8**	**61.2**	**79.3**	**85.8**	**74.3**

Table 1 Comparisons between TUNA and baseline models on multimodal understanding benchmarks. Results with model size greater than 13B are grayed. **Bold**: best results among each section. Underline: second-best. * indicates the results based on our evaluation scripts.

Our image captioning objective aligns with the pretraining objectives of strong semantic encoders, such as SigLIP 2 (Tschannen et al., 2025) and the Qwen2.5-VL (Bai et al., 2025) vision encoder. Image captioning has also been shown to provide semantic richness comparable to contrastive learning (Tschannen et al., 2023), thereby enhancing our unified representation's visual understanding capability. Meanwhile, the text-to-image generation objective trains the flow matching head to generate images from text conditions, laying the groundwork for later image editing and text-to-video generation tasks. Additionally, this objective allows generation gradients to flow back into the representation encoder, further aligning our unified visual representation with both understanding and generation tasks.

Stage 2: full model continue pretraining. In the second training stage, we unfreeze the LLM decoder and pretrain the entire model using the same image captioning and text-to-image generation objectives from Stage 1. During later training steps of Stage 2, we further introduce image instruction-following, image editing, and video-captioning datasets to extend the model's capabilities. This stage enables TUNA to perform more complex multimodal reasoning and generation tasks, bridging the gap between basic visual-text alignment and higher-level instruction-driven multimodal understanding and generation.

Stage 3: supervised finetuning (SFT). Finally, in the third stage, we conduct supervised fine-tuning (SFT) using a combination of image editing, image/video instruction-following, and high-quality image/video generation datasets, trained with a reduced learning rate. This stage further refines TUNA 's capabilities, improving its performance and generalization across diverse multimodal understanding and generation tasks.

3 Experiments

3.1 Experiment Setup

Implementation details. We verify TUNA with two LLM models at different scales, *i.e.*, Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Bai et al., 2025). In the pretraining stage, we optimize the representation encoder, projection layers, and diffusion head with AdamW (Loshchilov and Hutter, 2017) using a learning rate of 1×10^{-4}. We train on images with a base resolution of 512×512, as well as alternative aspect ratios that yield a similar number of visual tokens. In the second stage, we enable end-to-end training after a linear warm-up

Models	Size	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
Generation-only Models								
SD3-Medium (Esser et al., 2024)	2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1 [Dev]† (Batifol et al., 2025)	12B	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Composite UMMs								
MetaQuery-XL† (Pan et al., 2025)	7B	-	-	-	-	-	-	0.80
Tar (Han et al., 2025)	7B	0.99	0.92	0.83	0.85	0.80	0.65	0.84
BLIP3-o (Chen et al., 2025a)	8B	-	-	-	-	-	-	0.84
UniWorld-V1† (Lin et al., 2025a)	12B	0.98	0.93	0.81	0.89	0.74	0.71	0.84
OmniGen2† (Wu et al., 2025c)	7B	0.99	0.96	0.74	0.98	0.71	0.75	0.86
1.5B-scale Native UMMs								
D-DiT (Li et al., 2025c)	2B	0.97	0.80	0.54	0.76	0.32	0.50	0.65
Show-o (Xie et al., 2024b)	1.5B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Janus-Pro (Chen et al., 2025b)	1.5B	0.98	0.82	0.51	0.89	0.65	0.56	0.73
Show-o2 (Xie et al., 2025a)	1.5B	0.99	0.86	0.55	0.86	0.46	0.63	0.73
Harmon (Wu et al., 2025d)	1.5B	0.99	0.86	0.66	0.85	0.74	0.48	0.76
TUNA	1.5B	**1.00**	**0.94**	**0.83**	**0.91**	**0.81**	**0.79**	**0.88**
7B-scale Native UMMs								
MUSE-VL (Xie et al., 2025b)	7B	-	-	-	-	-	-	0.57
Transfusion (Zhou et al., 2024)	7B	-	-	-	-	-	-	0.63
Emu3 (Wang et al., 2024c)	8B	-	-	-	-	-	-	0.66
Show-o2 (Xie et al., 2025a)	7B	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76
Janus-Pro (Chen et al., 2025b)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
BAGEL† (Deng et al., 2025a)	14B	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88
Mogao (Liao et al., 2025)	7B	**1.00**	**0.97**	0.83	0.93	0.84	0.80	0.89
TUNA	7B	**1.00**	**0.97**	0.81	0.91	**0.88**	**0.83**	**0.90**

Table 2 Image generation results on GenEval. † refers to methods using LLM rewriters. **Bold**: best results among each section. Underline: second-best.

of 2K steps and continue optimization with the same learning rate. At this point, we extend the training data to include video-caption pairs and editing data. In the final stage, we perform supervised fine-tuning for instruction following on our curated SFT corpus with a smaller learning rate of 2×10^{-5}. Due to the substantial computational cost of video training, the 7B variant is trained without video data.

3.2 Main Results

Image understanding. We evaluate TUNA's multimodal understanding capabilities on nine benchmarks, including general VQA benchmarks such as MME (Fu et al., 2025a), GQA (Hudson and Manning, 2019), RealWorldQA (xAI) and SEED-Bench (Li et al., 2023b); knowledge-intensive benchmarks such as MMMU (Yue et al., 2024), MMStar (Chen et al., 2024b), and AI2D (Kembhavi et al., 2016); and text-centric benchmarks including ChartQA (Masry et al., 2022) and OCRBench (Liu et al., 2024b). As shown in Table 1, both 1.5B and 7B TUNA achieve state-of-the-art results across nearly all benchmarks, demonstrating strong and consistent performance. Notably, TUNA delivers competitive image understanding results compared to understanding-only models and outperforms many composite UMMs and UMMs with larger model sizes, highlighting the effectiveness of its unified representations.

Image generation. We evaluate TUNA's image generation performance on three benchmarks: GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024) and OneIG-Bench (Chang et al., 2025). Results are presented in Table 2 and Table 3. Across all three benchmarks, TUNA consistently outperforms contemporary approaches such as Janus-Pro, BAGEL and Mogao, achieving state-of-the-art results for both the 1.5B and 7B variants. Notably, TUNA shows a substantial advantage in text rendering quality in OneIG-Bench, indicating its strong semantic understanding capability when generating images from complex instructions containing visual text-related information. Our results show that TUNA consistently outperforms models with decoupled visual representations on image generation tasks, underscoring the strength and robustness of its unified representation design.

Models	Size	DPG-Bench						OneIG-Bench					
		Global	Entity	Attribute	Relation	Other	Overall	Alignment	Text	Reasoning	Style	Diversity	Average
Generation-only Models													
FLUX.1 [Dev] (Batifol et al., 2025)	12B	82.10	89.50	88.70	91.10	89.40	84.00	0.79	0.52	0.25	0.37	0.24	0.43
Qwen-Image (Wu et al., 2025a)	20B	91.32	91.56	92.02	94.31	92.73	88.32	0.88	0.89	0.31	0.42	0.20	0.54
1.5B-scale Native UMMs													
Show-o (Xie et al., 2024b)	1.3B	-	-	-	-	-	-	0.70	0.00	0.21	**0.36**	**0.24**	0.25
Show-o2 (Xie et al., 2025a)	1.5B	87.53	**90.38**	91.34	90.30	**91.21**	85.02	0.80	0.13	**0.27**	0.35	0.19	0.35
TUNA	1.5B	**88.87**	90.32	**91.71**	**91.79**	90.14	**86.03**	**0.82**	**0.77**	0.25	**0.36**	0.20	**0.48**
7B-scale Native UMMs													
Emu3-DPO (Wang et al., 2024c)	8B	-	-	-	-	-	81.60	-	-	-	-	-	-
Janus-Pro (Chen et al., 2025b)	7B	86.90	88.90	89.40	89.32	89.48	84.19	0.55	0.00	0.14	0.28	**0.37**	0.27
Mogao (Liao et al., 2025)	7B	82.37	90.03	88.26	**93.18**	85.40	84.33	-	-	-	-	-	-
BAGEL (Deng et al., 2025a)	14B	88.94	90.37	**91.29**	90.82	88.67	85.07	0.77	0.24	0.17	0.37	0.25	0.36
Show-o2 (Xie et al., 2025a)	7B	89.00	**91.78**	89.96	91.81	**91.64**	86.14	0.82	0.00	0.23	0.32	0.18	0.31
TUNA	7B	**90.42**	91.68	90.94	91.87	90.73	**86.76**	**0.84**	**0.82**	**0.27**	**0.40**	0.19	**0.50**

Table 3 Image generation results on DPG-Bench. **Bold**: best results among each section. Underline: second-best.

Models	Size	ImgEdit-Bench										GEdit-Bench		
		Add	Adj.	Ext.	Rep.	Rm.	Bg.	Sty.	Hyb.	Act.	Overall	G-SC	G-PQ	G-Overall
Generation-only Models														
FLUX.1 Kontext [Pro] (Batifol et al., 2025)	12B	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00	7.02	7.60	6.56
Qwen-Image (Wu et al., 2025a)	20B	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27	8.00	7.86	7.56
Native or Composite UMMs														
OmniGen (Xiao et al., 2025)	3.8B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96	5.96	5.89	5.06
BAGEL (Deng et al., 2025a)	14B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20	7.36	6.83	6.52
UniWorld-V1 (Lin et al., 2025a)	12B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26	4.93	7.43	4.85
OmniGen2 (Wu et al., 2025c)	4B	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	4.68	3.44	7.16	6.77	6.41
TUNA	7B	**4.46**	**4.52**	**2.47**	**4.68**	**4.58**	**4.56**	4.73	**4.07**	**4.69**	**4.31**	**7.79**	**7.48**	**7.29**

Table 4 Image editing results on ImgEdit-Bench and GEdit-Bench. For ImgEdit-Bench, we test editing performance across various dimensions, including 'Add', 'Adjust', 'Extract', 'Replace', 'Remove', 'Background', 'Style', 'Hybrid', and 'Action'. For GEdit-Bench, "G-SC" and "G-PQ" denote "G-Semantic Consistency" and "G-Perceptual Quality", respectively. **Bold**: best results among each section. Underline: second-best.

Image editing. We employ ImgEdit-Bench (Ye et al., 2025) and GEdit-Bench as our evaluation suite for image editing. As shown in Table 4, TUNA achieves an overall score of 4.31 on ImgEdit-Bench, ranking highest among all UMMs. TUNA's performance is also comparable to generation-only models such as FLUX.1 Kontext (Batifol et al., 2025) and Qwen-Image (Wu et al., 2025a). For GEdit-Bench, although TUNA performs slightly below the best generation-only model (Qwen-Image (Wu et al., 2025a)), it again achieves the highest overall score among all unified models. TUNA's consistently strong results on both ImgEdit-Bench and GEdit-Bench demonstrate its robust image editing capability and highlight the effectiveness of our unified visual representation when handling visual generation tasks that require precise semantic understanding and accurate prompt following.

Video understanding. We employ four video understanding benchmarks to evaluate TUNA: MVBench (Li et al., 2024b), Video-MME (Fu et al., 2025b), LongVideoBench (Wu et al., 2024a) and LVBench (Wang et al., 2025b). As shown in Table 5, TUNA outperforms Show-o2 on MVBench and Video-MME, while achieving competitive results on LongVideoBench and LVBench. Notably, despite being only a 1.5B-parameter model, TUNA performs on par with larger understanding-only models on MVBench and LVBench, demonstrating the efficiency and effectiveness of our unified representation for video understanding tasks.

Video generation. We evaluate TUNA on VBench (Huang et al., 2024) for text-to-video generation, comparing it against other UMMs and generation-only models. As shown in Table 6, TUNA achieves state-of-the-art performance, surpassing all existing UMMs capable of video generation, while using only a 1.5B-parameter LLM decoder. This demonstrates the efficiency and scalability of our unified architecture for high-quality video generation.

Models	Size	#Frames	MVBench	Video-MME	LongVideoBench	LVBench
			test	w/o sub	val	test
Understanding-only Models (LMMs)						
GPT-4o (OpenAI, 2024)	-	-	-	71.9	66.7	48.9
Gemini-1.5-Pro (Team et al., 2024)	-	-	54.2	75.0	64.0	33.1
LongVA (Zhang et al., 2024a)	7B	64	49.2	52.6	51.8	-
VideoLLaMA2 (Cheng et al., 2024)	7B	16	54.6	47.9	-	-
LLaVA-OV (Li et al., 2024a)	7B	32	56.7	58.2	56.5	26.9
1.5B-scale Native UMMs						
Show-o2 (Xie et al., 2025a)	1.5B	32	49.8	48.0	49.2	-
TUNA	1.5B	49	**54.4**	**49.1**	**49.7**	**27.4**

Table 5 Experimental results on video understanding benchmarks. #Frames denotes the number of frames used during inference. **Bold**: best results. Underline: second-best.

Models	Size	QS	SS	SC	BC	TF	MS	DD	AQ	IQ	OC	MO	HA	C	SR	S	AS	TS	OC'	Total
Generation-only Models																				
CogVideoX	5B	82.75	77.04	96.23	96.52	98.66	96.92	70.97	61.98	62.90	85.23	62.11	99.40	82.81	66.35	53.20	24.91	25.38	27.59	81.61
Native or Composite UMMs																				
VILA-U	7B	76.26	65.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	74.01
HaploOmni	7B	-	-	96.40	97.60	-	96.80	65.30	-	-	-	-	-	-	-	34.60	-	-	-	78.10
Emu3	8B	-	-	95.32	**97.69**	-	**98.93**	**79.27**	59.64	-	86.17	44.64	77.71	-	68.73	37.11	20.92	-	-	80.96
Show-o2	1.5B	82.10	78.31	**97.28**	96.78	97.68	98.25	40.83	65.15	**67.06**	94.81	76.01	95.20	80.89	62.61	57.67	**23.29**	**25.27**	27.00	81.34
TUNA	1.5B	**84.32**	**83.04**	95.99	96.72	**98.02**	98.33	69.39	**65.88**	66.83	**95.41**	**92.31**	**97.50**	**87.67**	**78.12**	**58.59**	23.18	24.68	**27.71**	**84.06**

Table 6 Video generation results on VBench. Full column names: QS: Quality Score, SS: Semantic Score, SC: Subject Consistency, BC: Background Consistency, TF: Temporal Flickering, MS: Motion Smoothness, DD: Dynamic Degree, AQ: Aesthetic Quality, IQ: Imaging Quality, OC: Object Class, MO: Multiple Objects, HA: Human Action, C: Color, SR: Spatial Relationship, S: Scene, AS: Appearance style, TS: Temporal Style, OC': Overall Consistency. **Bold**: best results among each section. Underline: second-best.

3.3 Ablation: Visual Representation Design

In this section, we conduct a series of ablation experiments to systematically assess the effectiveness of our model architecture and training pipeline. For all experiments, we use a lightweight variant of TUNA built on the Qwen2.5-1.5B LLM with a smaller flow matching head. Using this setup, we evaluate three visual representation designs:

1. Decoupled representations, using SigLIP 2 features for understanding and Wan 2.2 VAE latents for generation (denoted as "Decoupled").
2. Show-o2-style unified representations, using a dual-path late fusion strategy to obtain the final representations (denoted as "Show-o2"). A detailed explanation of this design can be found in Section 3.4.
3. TUNA's unified representation, initialized from three different pretrained representation encoders: SigLIP (Zhai et al., 2023), SigLIP 2 (Tschannen et al., 2025), and DINOv3[3] (Siméoni et al., 2025).

All models are trained on a subset of our training data using a two-stage training pipeline (corresponding to Stage 1 and Stage 3 in Section 2.3), with an equal number of training steps per stage. Our ablation study results are shown in Table 7.

Unified vs. decoupled visual representation. Comparing Model 8 and Model 12 in Table 7, we observe that our unified representation consistently outperforms the decoupled setting across all understanding and generation benchmarks. Comparing Models 2 and 5 with Model 8, we find that training a unified model using decoupled visual representations results in substantial degradation on understanding tasks, compared to only training the model on understanding data. In contrast, Model 12 surpasses Model 3 on most understanding benchmarks and outperforms Model 6 across all generation benchmarks. These results indicate that our unified representation suffers far less from representation conflicts than decoupled designs, enabling stronger performance in both understanding and generation.

[3]Model code: `siglip-so400m-patch14-384`, `siglip2-so400m-patch16-512` and `dinov3-vith16plus-pretrain-lvd1689m`.

Models	ID	Data	Understanding				Generation	
			MME-p	MMMU	SEED	GQA	GenEval	DPG
Show-o2 (Wan 2.1 VAE + SigLIP)	1	Und. Only	1351	36.1	62.1	56.8	-	-
Decoupled (SigLIP 2 only)	2		1392	38.2	62.9	58.1	-	-
TUNA (Wan 2.2 VAE + SigLIP 2)	3		1386	37.6	62.9	57.4	-	-
Show-o2 (Wan 2.1 VAE + SigLIP)	4	Gen. Only	-	-	-	-	76.2	82.56
Decoupled (Wan 2.2 VAE only)	5		-	-	-	-	77.3	82.87
TUNA (Wan 2.2 VAE + SigLIP 2)	6		-	-	-	-	77.8	83.33
Show-o2 (Wan 2.1 VAE + SigLIP)	7	Und. & Gen.	1339	35.4	61.7	55.9	75.9	82.32
Decoupled (Wan 2.2 VAE + SigLIP 2)	8		1346	37.2	61.4	56.5	78.3	83.50
TUNA (Wan 2.1 VAE + SigLIP)	9		1358	35.9	64.2	57.2	77.2	83.29
TUNA (Wan 2.2 VAE + SigLIP)	10		1349	36.3	64.6	57.4	76.9	83.10
TUNA (Wan 2.1 VAE + SigLIP 2)	11		1379	37.7	65.9	58.4	79.1	83.98
TUNA (Wan 2.2 VAE + SigLIP 2)	12		1361	38.1	66.5	58.2	79.4	84.20
TUNA (Wan 2.2 VAE + DINOv3)	13		1396	37.3	65.6	58.6	78.9	84.08

Table 7 Ablation study results. "Und. Only", "Gen. Only" and "Und. & Gen." refer to models trained with understanding data only, generation data only, and both data, respectively.

Selection of representation encoders. We find that TUNA's unified representation generally benefits from stronger representation encoders. As shown in Table 7, comparing Models 10, 12, and 13, both SigLIP 2 (400M parameters) and DINOv3 (800M) outperform SigLIP (400M) on all benchmarks. Furthermore, comparing Model 9 to Model 11 and Model 10 to Model 12, we observe that regardless of which VAE encoder is used in the model, replacing SigLIP with SigLIP 2 in the representation encoder consistently improves performance across all understanding and generation benchmarks. We ultimately adopt SigLIP 2 for TUNA as it delivers comparable understanding performance, superior generation quality relative to DINOv3, and maintains a significantly smaller model size.

Understanding-generation synergy. Our experimental results on training models exclusively on either understanding (Models 1, 2 and 3) or generation data (Models 4, 5 and 6) demonstrate that TUNA benefits from joint training on both data types. Specifically, we observe that Model 12 surpasses Model 3 on understanding benchmarks and Model 6 on generation benchmarks. Although the comparison between Model 2 and Model 3 shows that TUNA's VAE + representation encoder architecture incurs a slight performance drop relative to using only a representation encoder (the standard setup for understanding-only models), our joint understanding + generation training pipeline largely compensates for this degradation. Specifically, Model 12 recovers its understanding performance and becomes comparable to or even better than Model 2 on several understanding benchmarks. Moreover, Model 12 substantially outperforms both Model 5 and Model 6 on all generation benchmarks. These results demonstrate the mutual enhancement between understanding and generation made possible by our unified visual representation design.

Comparison with Show-o2. A closely related work to ours is Show-o2 (Xie et al., 2025a), which uses a dual-path late-fusion mechanism, merging features from separate VAE and semantic encoders via a fusion layer to build unified representations. In contrast, TUNA extracts unified representations directly from VAE latents using a semantic encoder, achieving deep feature fusion across all layers in the semantic encoder. Comparing Model 7 with Models 9, 10, 11 and 12 in Table 7, our unified representation consistently outperforms Show-o2 on all benchmarks, regardless of the choice of the VAE encoder and the representation encoder. Model 1 vs. 3 and Model 4 vs. 6 further show that Show-o2 underperforms even when trained on a single task. We attribute this to its late-fusion strategy, which introduces representation conflicts and degrades overall performance.

3.4 Discussion: Unified Representation Analysis

As discussed in Section 3.3, both TUNA and Show-o2 (Xie et al., 2025a) employ unified visual representations for understanding and generation, but they construct these representations in fundamentally different ways. In this section, we first describe Show-o2's unified visual representation design in detail, and then provide an in-depth analysis of why TUNA's unified representation achieves superior performance than Show-o2.

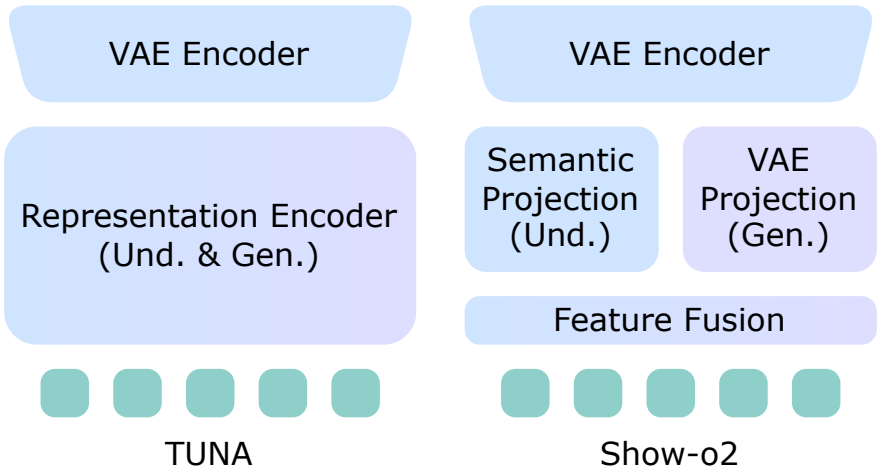


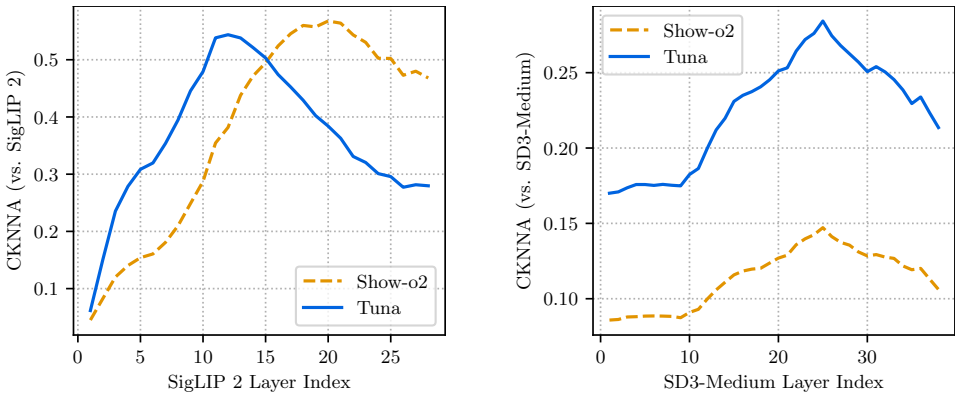
Figure 4 Comparison between TUNA and Show-o2 on how unified visual representations are produced.

As illustrated in Figure 4, Show-o2 constructs unified visual representations using a dual-path feature fusion mechanism. The input image or video is first encoded by a VAE encoder, after which the latent is processed through two parallel branches. The semantic projection branch feeds the VAE latents into a set of semantic layers to extract features for understanding tasks. The VAE projection branch applies 2D patch embedding layers to produce features tailored for generation tasks. Importantly, the semantic layers are *pre-distilled* using a frozen representation encoder: given the same image, their outputs are first aligned with a pretrained SigLIP model before conducting end-to-end training of the Show-o2 model. This pre-distillation stage is proposed to preserve semantic understanding capability. Finally, Show-o2 merges the outputs of both paths using a feature fusion layer to obtain its unified visual representation.

To better understand why our unified representation yields superior performance, we perform a representation alignment analysis using CKNNA scores (Huh et al., 2024) with respect to two reference models: (1) a strong semantic encoder SigLIP 2 (Tschannen et al., 2025), and (2) a strong generation model SD3-Medium (Esser et al., 2024). Concretely, we extract unified visual representations from TUNA and Show-o2 based on 1,024 images from the Wikipedia Captions dataset (Srinivasan et al., 2021) and compute their CKNNA scores relative to features from all intermediate layers of the two reference models. The results are presented in Figure 5a and Figure 5b.

As shown in the figures, both TUNA and Show-o2 exhibit strong alignment with the SigLIP 2 intermediate features, with CKNNA scores exceeding 0.5. This high similarity reflects their strong semantic understanding capability, consistent with their solid performance on multimodal understanding tasks. On the other hand, TUNA's unified representation shows consistently higher alignment with the SD3-Medium intermediate features compared to Show-o2, indicating that TUNA learns a more balanced unified representation suited for both understanding and generation. In contrast, Show-o2 remains biased toward semantic features, which limits its generation quality.

The above findings prompt us to further investigate why Show-o2's dual-path fusion mechanism produces biased features toward semantic understanding. To analyze this, we compute CKNNA scores between Show-o2's final fused features and the intermediate features from its understanding (semantic projection) and generation (VAE projection) branches before fusion. We find that Show-o2's unified representation exhibits a strong correlation with its understanding branch (CKNNA=0.45) but a very weak correlation with the generation branch (CKNNA=0.07). This demonstrates that the late-fusion strategy merges features in an imbalanced manner, causing the representation to remain dominated by semantic information. In contrast, TUNA's end-to-end training of the unified representation on both objectives enables early fusion of understanding and generation signals at every layer of the representation encoder. This layer-wise interaction captures richer cross-task dependencies and is inherently more robust than the late-fusion strategy adopted in Show-o2.



(a) Alignment to SigLIP 2 **(b)** Alignment to SD3-Medium

Figure 5 Representation alignment analysis with SigLIP 2 and SD3-Medium. For both TUNA and Show-o2, we extract visual representations at the input layer of the LLM decoder.

3.5 Qualitative Results

Image generation. We compare TUNA with state-of-the-art generation-only and unified models across diverse image generation instructions in Figure 6. In the first two examples, TUNA exhibits strong text rendering ability, accurately reproducing all visual text in the prompts without errors. In the whiteboard example, TUNA is the only model that correctly places an underline beneath "with everyone", demonstrating precise prompt-following capability. Moreover, TUNA accurately generates two black shelves, one containing books and markers on top and the other containing black cloth and hand sanitizer at the bottom, each in the correct position. Other models either fail to produce the correct number of shelves or place the wrong items on them. These results show that TUNA excels at compositional image generation, enabled by its unified visual representation with strong semantic understanding capabilities. In the "tuna" example, both TUNA and Flux (Batifol et al., 2025) successfully render the Hawaiian shirt, while other models either fail to depict the shirt or generate an incorrect tuna body. Finally, in the "red t-shirt" example, TUNA accurately reflects the "classic 1960s Walt Disney animation style" and correctly includes all required elements from the prompt, maintaining a coherent and well-structured composition.

Image editing. We compare TUNA with BAGEL (Deng et al., 2025a), Qwen-Image (Wu et al., 2025a), and Flux.1 Kontext (Batifol et al., 2025) on image editing tasks in Figure 7. As shown, TUNA not only correctly performs explicit editing operations, such as style transfer (photorealistic → hand-sculpted claymation in the "dog" example), environment change (daylight → nighttime in the "red car" example), and object replacement (boat → puppy with a swim ring in the "boat" example), but also handles more implicit and nuanced instructions, such as applying lighting from the left side in the "doll" example. These results further highlight TUNA's strong semantic understanding and high-fidelity image generation capabilities.

Video generation. We present TUNA's video generation results in Figure 8. The model produces high-fidelity videos across a wide range of instructions, demonstrating the strength of its unified visual representation space for jointly modeling both images and videos.

4 Related Work

4.1 Large Multimodal Models

Large multimodal models (LMMs) aim to generate text responses from multimodal inputs spanning images, videos, and text. Early LMMs such as Flamingo (Alayrac et al., 2022) and Idefics (Laurençon et al., 2023) introduced cross-attention layers to enable interaction between visual and linguistic features. Modern LMMs generally follow the LLaVA paradigm (Liu et al., 2023a), where visual inputs are encoded by a vision encoder (e.g., CLIP (Radford et al., 2021)) and then concatenated with text tokens for joint processing by a language model decoder. Recent research advances focus on improving instruction-following through higher-quality training data (Liu et al., 2024a; Li et al., 2024a; Chen et al., 2024a; Li et al., 2024b; Ren et al., 2024; Zhang

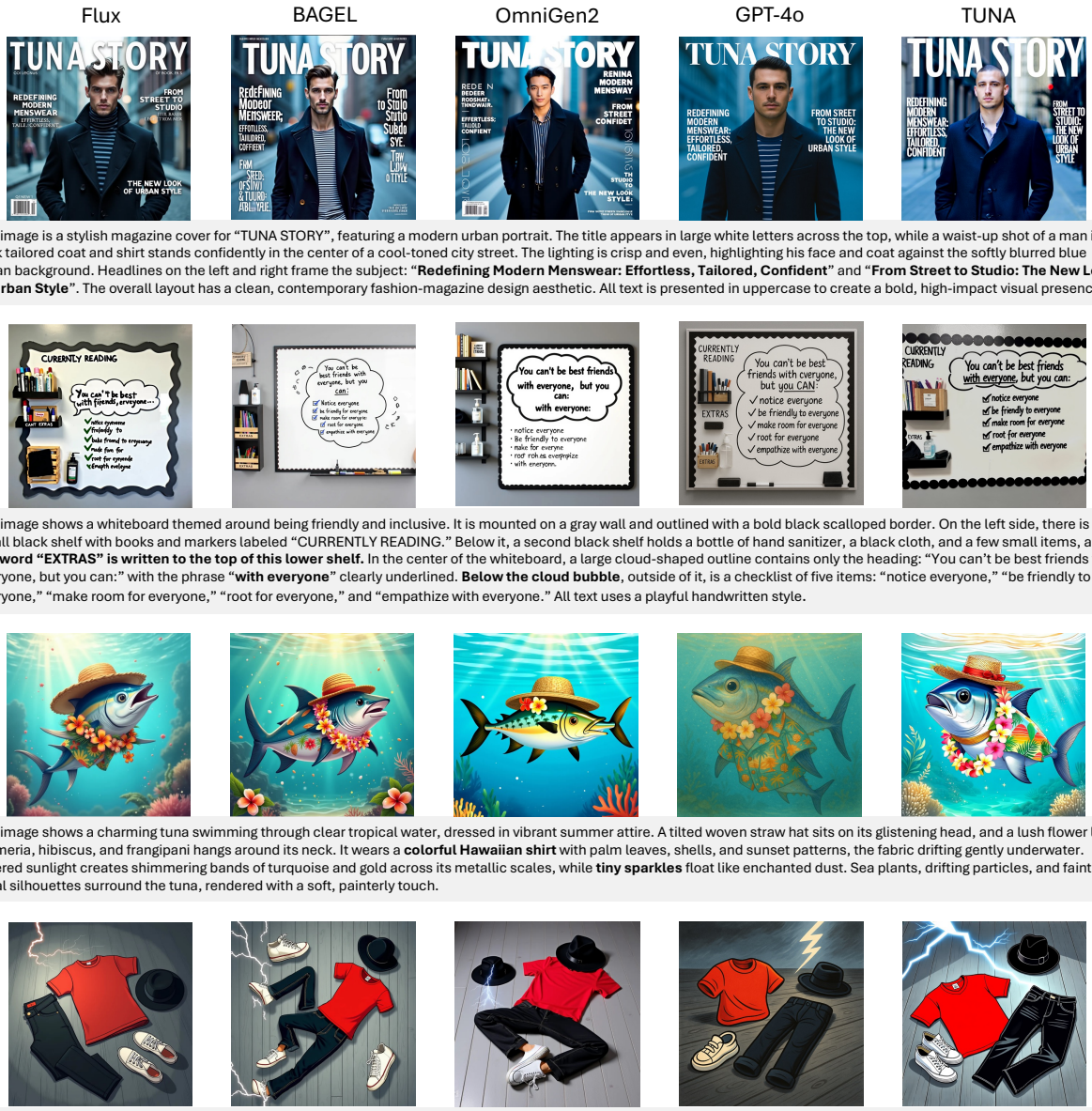


Figure 6 Qualitative comparison between TUNA and baseline models on image generation tasks. The instructions that are correctly reflected in our results but failed in some of the baseline models are **bolded**.

et al., 2024b; Wiedmann et al., 2025; An et al., 2025), developing stronger vision encoders capable of handling higher-resolution images (Liu et al., 2024a; Laurençon et al., 2024; Wang et al., 2024b; Bai et al., 2025), extending LMMs to interleaved image (Laurençon et al., 2024; Li et al., 2024a; Jiang et al., 2024) and video understanding (Maaz et al., 2023; Lin et al., 2023; Zhang et al., 2024a; Li et al., 2024b,d; Ren et al., 2025), and incorporating reinforcement learning with thinking modes (Deng et al., 2025b; Huang et al., 2025; Feng et al., 2025) or pixel-space reasoning (Su et al., 2025a,b; Liu et al., 2025a).

4.2 Diffusion Generative Models

Diffusion generative models have become the de facto backbone of high-fidelity image (Esser et al., 2024; Batifol et al., 2025; Li et al., 2024e; Wu et al., 2025a; Liu et al., 2023b,c, 2025b) and video (Kong et al., 2024; Seawead et al., 2025; Wan et al., 2025; Liu et al., 2025c) synthesis. Modern large-scale visual generation

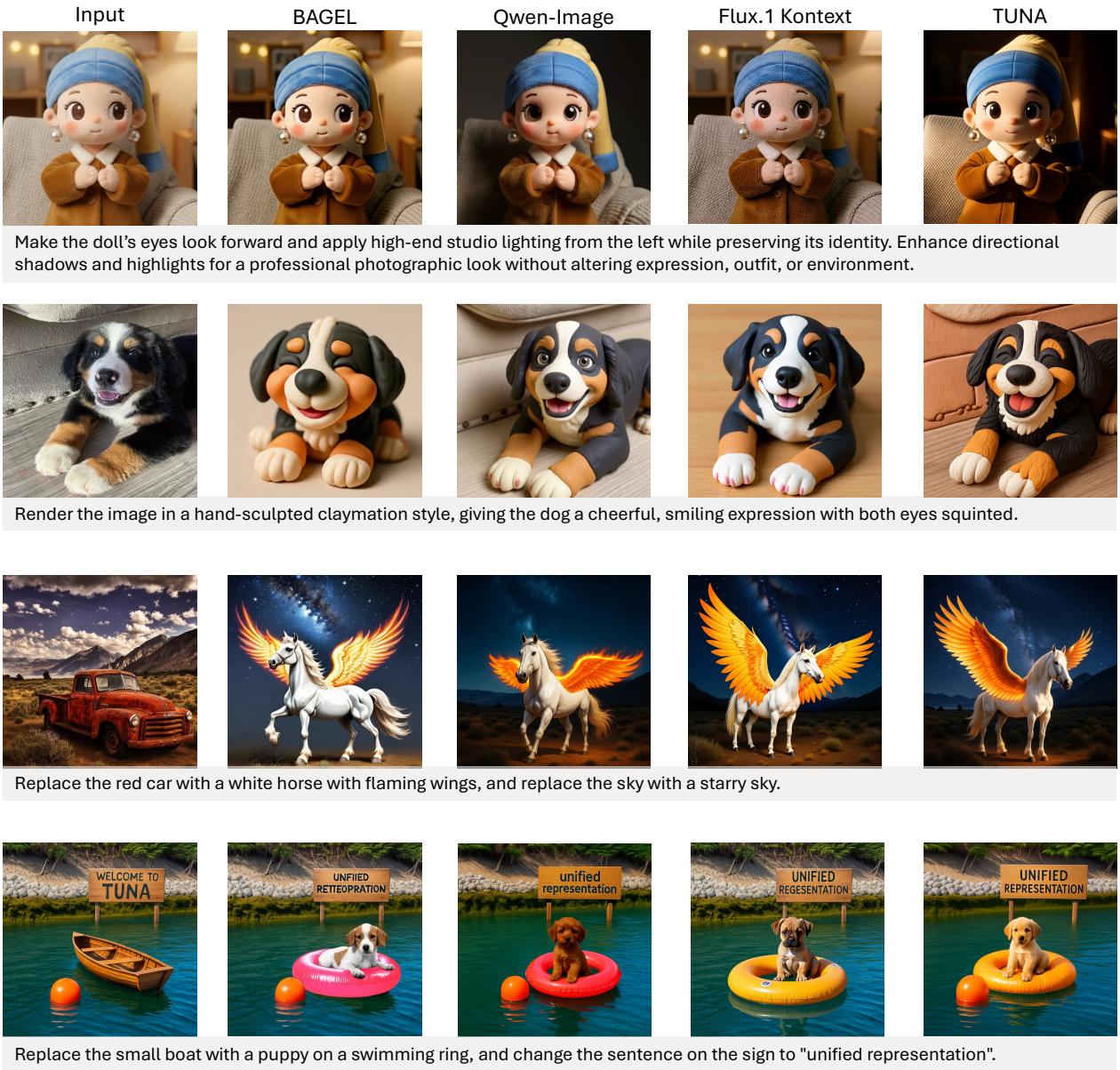


Figure 7 Qualitative comparison between TUNA and baseline models on image editing tasks.

models typically apply diffusion in a continuous latent space defined by a learned VAE, following the Latent Diffusion Model (LDM) paradigm (Rombach et al., 2022), which offers superior perceptual quality and sampling efficiency compared to autoregressive decoding of long sequences of discrete tokens based on VQ-VAE (Van Den Oord et al., 2017; Esser et al., 2021). Within diffusion itself, latent-space models (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024) are generally preferred over pixel-space approaches (Dhariwal and Nichol, 2021; Saharia et al., 2022) because they reduce computational cost, ease scaling to higher resolutions, and allow the denoising network to focus on semantically meaningful structure rather than low-level pixel noise. Architecturally, diffusion backbones have evolved from convolutional U-Net designs (Ronneberger et al., 2015; Ho et al., 2020) to diffusion transformers (DiT) (Peebles and Xie, 2023; Ma et al., 2024); In parallel, the learning objective has been generalized from Gaussian noise prediction and score matching (Ho et al., 2020; Song et al., 2020) to more expressive formulations such as rectified flows (Liu et al., 2022) and flow matching objectives (Lipman et al., 2022; Albergo et al., 2023).

A bird building a nest from twigs and leaves.

An elderly woman with short, curly white hair and a beige sweater sits on a gray couch, holding a light blue mug in her right hand. She rests her left arm on a green pillow, smiling as she looks off to the right. Behind her, a white shelf holds a glass vase with a gray candle, a stack of books, and a wicker basket. A large green plant stands to the left, and the background is a plain white wall. The woman's gaze shifts to the camera, and she raises the mug to her lips, taking a sip. She then sets the mug down on the couch beside her, her eyes never leaving the camera. The scene is well-lit, with soft, natural light illuminating the room.

A person stands at the edge of an infinity pool, arms outstretched, gazing out at a vibrant sunset over the ocean. The individual is silhouetted against the colorful sky, wearing a long-sleeved shirt and shorts. The pool's edge is barely visible, with ripples on the water's surface reflecting the sky's hues. In the background, a small island with palm trees and rocks meets the horizon, where the sun sets over the calm ocean waters. The sky transitions from blue to orange, yellow, and pink, with scattered clouds. The scene exudes a serene, peaceful atmosphere, capturing a moment of contemplation or joy at the natural beauty of the surroundings.

A majestic male lion with a thick, dark-brown mane and golden-brown fur lies on the ground, facing right. His front paws are stretched out in front of him, and his head is slightly raised, gazing into the distance. The lion's mane is long and shaggy, framing his face and neck. In the background, a tree trunk and some dry grass are visible, along with a large rock formation to the right. The ground is covered in light-brown dirt, and the overall atmosphere suggests a sunny day in a savannah or zoo enclosure. The lion's expression is calm and regal, exuding a sense of power and serenity.

The video shows a single red balloon drifting gently forward, its string hanging loosely as it floats at a steady, unhurried pace. The balloon rises and dips slightly with the movement of the air, shifting left or right in small, natural sways as it continues down the street. The scene unfolds along a old-town street treated as a full visual object rather than simple background: narrow cobblestone paths, tall stone buildings with ornate window frames, iron balconies, and warm-toned façades lining both sides. Soft afternoon light casts long shadows across café tables, bicycles, and shop signs, all subtly passing by as the balloon moves. The architecture's depth, the texture of the street, and the layered rows of buildings create a clear sense of space, framing the slow, floating journey of the red balloon through the historic European street.

Figure 8 Qualitative results for TUNA on the task of text-to-video generation.

4.3 Unified Multimodal Models

Unified multimodal models (UMMs) have gained growing attention for their ability to flexibly generate text and visual content from diverse multimodal inputs. Recent approaches (Luo et al., 2025) such as MetaQuery (Pan et al., 2025), BLIP-3o (Chen et al., 2025a), and UniWorld-V1 (Lin et al., 2025a) achieve this by connecting understanding-only and generation-only models through learnable adapters. While achieving promising results, their capabilities largely rely on pretrained task-specific models, limiting the potential synergy between understanding and generation. In contrast, native UMMs are pretrained from scratch to perform both tasks within a single unified architecture. Among these works, models such as the Janus series (Wu et al., 2025b; Ma et al., 2025c; Chen et al., 2025b) and UniFluid (Fan et al., 2025) adopt decoupled visual representations for understanding and generation. BAGEL (Deng et al., 2025a), Mogao (Liao et al., 2025), and OneCAT (Li et al., 2025a) further use MoE-style architectures to route inputs separately, mitigating conflicts between distinct representations from the decoupled vision encoders. Alternatively, models like Chameleon (Team, 2024), Transfusion (Zhou et al., 2024), Harmon (Wu et al., 2025d), and the Show-o series (Xie et al., 2024b, 2025a) employ unified visual representations for both tasks. While being more efficient, these models often exhibit weaker or imbalanced performance, excelling in one task but underperforming in

the other. TUNA overcomes these limitations by learning balanced, unified visual representations and achieves strong performance in both understanding and generation.

4.4 Representation in Multimodal Models

Recent studies have explored learning better representations to enhance multimodal understanding and generation models. From the perspective of improving understanding models, methods such as Ross (Wang et al., 2024a), GenHancer (Ma et al., 2025b) and ASVR (Wang et al., 2025a) enhance multimodal understanding by introducing generation or reconstruction objectives, encouraging the model to capture fine-grained visual details. Conversely, to improve generative models, approaches such as REPA (Yu et al., 2024) and VA-VAE (Yao et al., 2025) align diffusion transformers or VAE representations with semantic vision encoders, thereby achieving stronger generative performance. Similarly, Dispersive Loss (Wang and He, 2025) introduces an auxiliary contrastive-like objective to further enhance generation quality.

In the domain of unified multimodal models, recent research has primarily focused on developing unified visual tokenizers that support both understanding and generation tasks. For instance, TokenFlow (Qu et al., 2025) and MUSE-VL (Xie et al., 2024c) adopt late-fusion strategies to merge features from separate understanding and generation encoders into quantized codebooks. DualToken (Song et al., 2025), UniTok (Ma et al., 2025a) and TokLIP (Lin et al., 2025b) train a single encoder to produce vector-quantized representations for both tasks. However, these methods rely on discrete representations, limiting their ability to perform high-fidelity visual generation. UniFlow (Yue et al., 2025) and UniLIP (Tang et al., 2025) adapt representation encoders into continuous unified visual tokenizers, but both rely on relatively complex alignment (*e.g.* self-distillation or reconstruction schemes). In contrast, TUNA learns a unified representation end-to-end under joint understanding and generation objectives, and is validated at a larger scale across more tasks. Moreover, UniLIP adopts a composite design where the unified features only serve as conditions for a separate pretrained generative model (SANA (Xie et al., 2024a)). On the other hand, TUNA trains a native unified model that jointly performs understanding and generation within a single framework.

5 Conclusion

We introduced TUNA, a native unified multimodal model that constructs a unified visual representation space by cascading a VAE encoder with a representation encoder. We train an LLM decoder and a flow matching head on this unified representation, achieving strong performance across image and video understanding, image and video generation, and image editing. TUNA not only surpasses prior UMM baselines but also performs competitively with leading understanding-only and generation-only models. Our ablation studies further show that (1) TUNA's unified representation space outperforms both Show-o2-style unified representations and decoupled representation designs, (2) stronger pretrained representation encoders consistently yield better performance within our framework, and (3) our unified visual representation design enables mutual enhancement between understanding and generation.

6 Acknowledgment

We would like to thank Yukang Yang (Princeton University), Ji Xie (UC Berkeley), and Jinheng Xie (NUS) for their constructive feedback on this project.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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.

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025.

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*, pages 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? *Advances in Neural Information Processing Systems*, 37:27056–27087, 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.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

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*, 2025a.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

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.

Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025a. https://arxiv.org/abs/2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025b.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025.

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

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.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 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.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

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

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023b.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025a.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29767–29779, 2025b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024b.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024c.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024d.

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*, 2024e.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2779–2790, 2025c.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

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*, 2025a.

Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation, 2025b. https://arxiv.org/abs/2505.05422.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024a. Accessed: 2025-02-14.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025a.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023b.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Customizable image synthesis with multiple subjects. *Advances in neural information processing systems*, 36:57500–57519, 2023c.

Zhiheng Liu, Ka Leong Cheng, Xi Chen, Jie Xiao, Hao Ouyang, Kai Zhu, Yu Liu, Yujun Shen, Qifeng Chen, and Ping Luo. Manganinja: Line art colorization with precise reference following. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5666–5677, 2025b.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception. *arXiv preprint arXiv:2508.15720*, 2025c.

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

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. *arXiv preprint arXiv:2510.13721*, 2025.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025a.

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

Shijie Ma, Yuying Ge, Teng Wang, Yuxin Guo, Yixiao Ge, and Ying Shan. Genhancer: Imperfect generative models are secretly strong vision-centric enhancers. *arXiv preprint arXiv:2503.19480*, 2025b.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025c.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

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

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

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.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025.

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*, pages 8748–8763. PmLR, 2021.

Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. *arXiv preprint arXiv:2412.00927*, 2024.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. *arXiv preprint arXiv:2503.11579*, 2025.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Wei Song, Yuran Wang, Zijia Song, Yadong Li, Haoze Sun, Weipeng Chen, Zenan Zhou, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

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

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025b.

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.

Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Wan-Video team. Wan: Open and advanced large-scale video generative models (wan2.2), 2025. https://github.com/Wan-Video/Wan2.2. GitHub repository, Apache-2.0 License.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi Wang. Autoregressive semantic visual reconstruction helps vlms understand better. *arXiv preprint arXiv:2506.09040*, 2025a.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024a.

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*, 2024b.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

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*, 2024c.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977, 2025b.

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*, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024a.

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*, 2025d.

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*, 2024b.

xAI. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v. Company news post.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024a.

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*, 2024b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025a.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024c.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24135–24146, 2025b.

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.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 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.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 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*, pages 9556–9567, 2024.

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, KunPeng Du, Yi Wang, Limin Wang, et al. Uniflow: A unified pixel flow tokenizer for visual understanding and generation. *arXiv preprint arXiv:2510.10575*, 2025.

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.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37:42672–42695, 2024.

Ran Zuo, Haoxiang Hu, Xiaoming Deng, Cangjun Gao, Zhengming Zhang, Yukun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenediff: Generative scene-level image retrieval with text and sketch using diffusion models. 2024.