This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

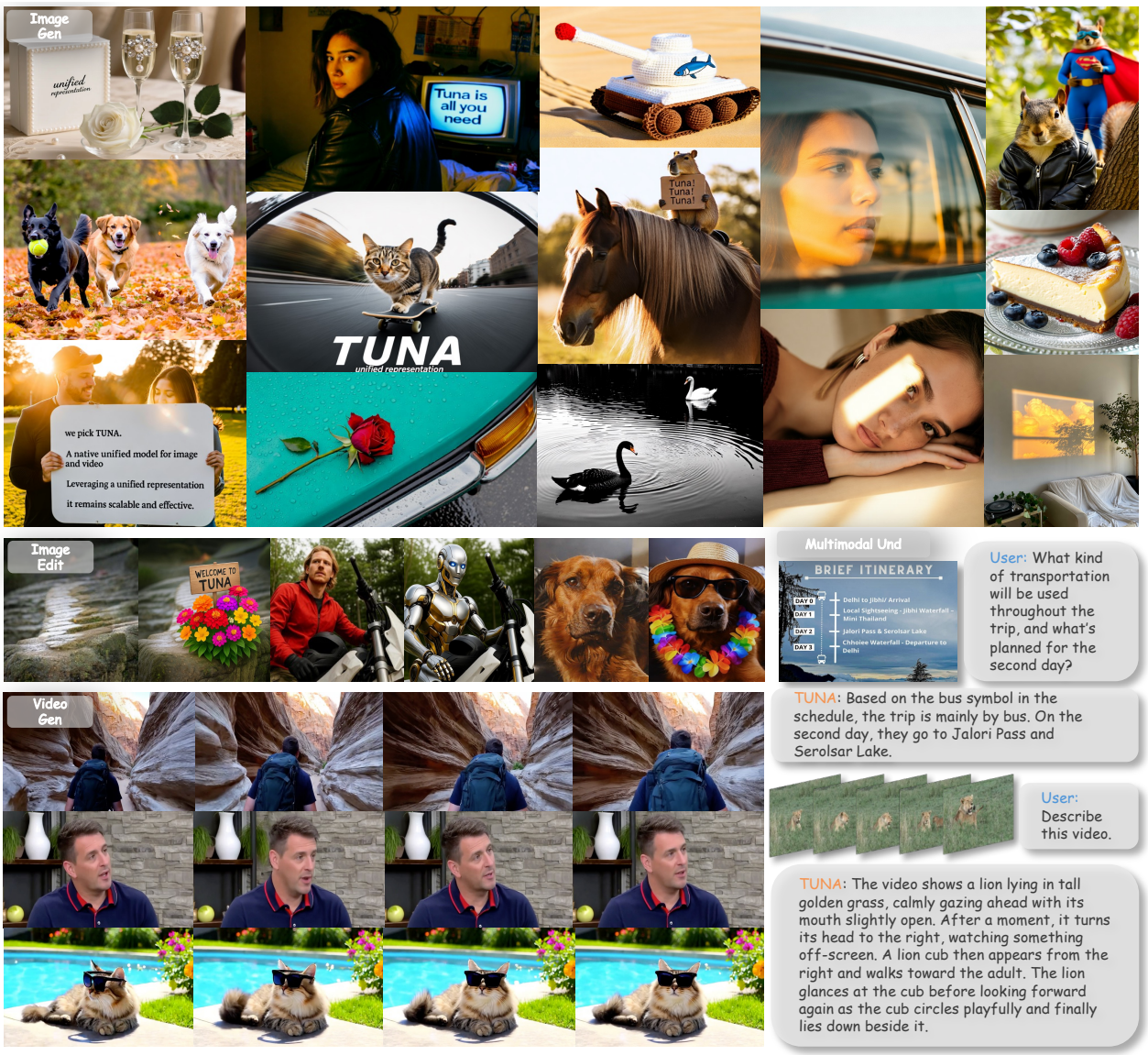


Figure 1. We present TUNA, a native unified multimodal model built on a unified visual representation, enabling diverse multimodal understanding and generation capabilities such as image and video understanding, image and video generation, and image editing.

Abstract

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.*

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 (UMMs) [7, 51, 71] 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 [7] and Mogao [28], 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 [80]) and a causal VAE encoder (*e.g.* Wan 2.1 VAE [55]) 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 recent studies like Ross [56] and REPA [77] 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 [51] and Transfusion [83] use VQ-VAE [8], while Harmon [64] utilizes the MAR encoder [26]. This unified design tends to favor one task at the expense of the other. Show-o2 [71] attempts to mitigate this issue by fusing SigLIP [80] and VAE [55] 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 [5] and 0.90 on GenEval [13]).

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,

[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.

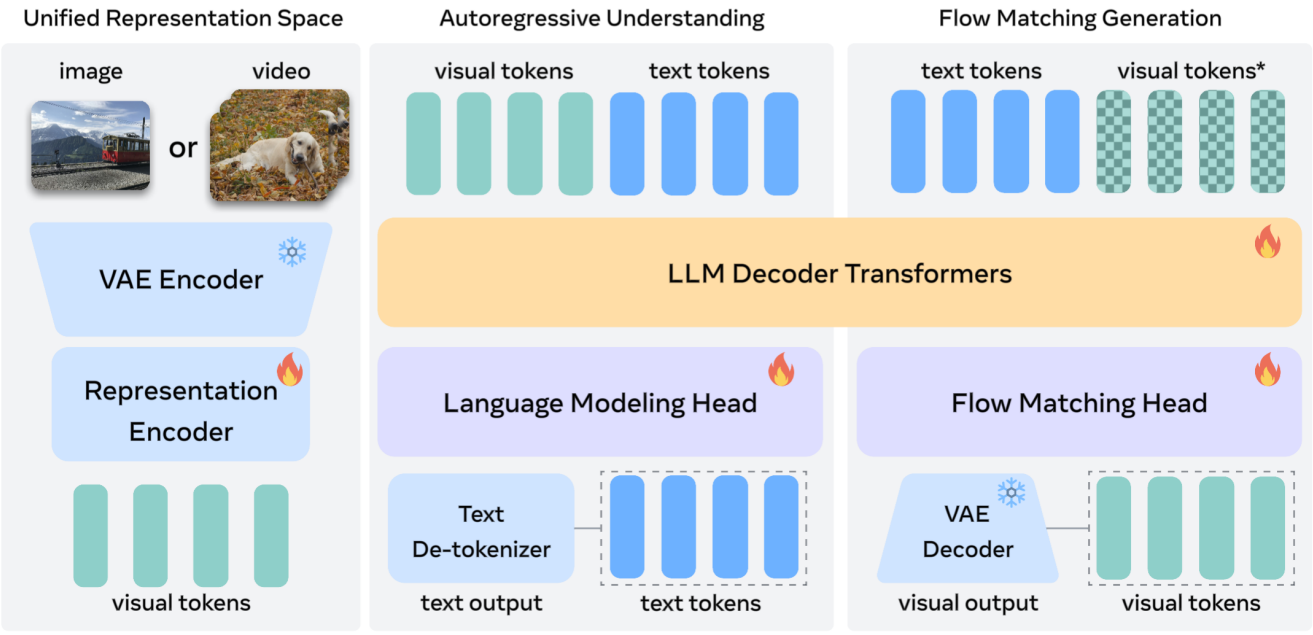


Figure 2. Overview of the TUNA architecture. Our model employs a VAE encoder and a representation encoder to construct unified visual representations, which are then combined with text tokens and processed by an LLM decoder. The decoder performs autoregressive text generation for understanding tasks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

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 [14, 49, 74] or diffusion models [3, 37, 55]. In practice, leading understanding-only models [2, 57] adopt autoregressive models for text generation. On the other hand, state-of-the-art image and video generators [9, 55] 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 [9, 55] outperform those using discrete representations [49], as discretization causes information loss and reduces fidelity. Similarly, multimodal understanding models [2, 57] typically rely on continuous semantic features (*e.g.*, CLIP [43] 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 [77] demonstrates that diffusion transformers benefit from aligning intermediate features with pretrained representation encoders like DINOv2 [39]. Concurrent to our work, RAE [82] 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 [34] and TokLIP [30] 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 [84], object recognition [21], and image retrieval [85].

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

[2]Due to higher generation fidelity and lower training costs, latent diffusion models are generally favoured over pixel diffusion [82].

- 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 [52], which downsamples the input by $16\times$ spatially and $4\times$ 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 SigLIP2 vision encoder Φ (patch size 16, pretrained resolution 512) to extract semantic features from the VAE latents. Since the VAE encoder has $16\times$ downsampling, we replace SigLIP2'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:

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

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

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

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 [2]) for joint multimodal processing. Following standard UMM practices [7, 70], 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 [71] and DiT [41]. For generation and editing tasks, we adopt multimodal 3D-RoPE [44, 48] over the concatenated text-visual sequence to handle interleaved instructions and visual content.

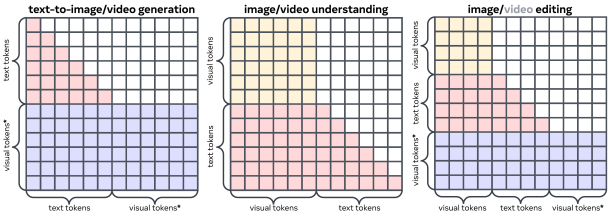


Figure 3. Attention masks in the LLM decoder for understanding and generation tasks. * indicates that the visual tokens are noised.

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.

Our image captioning objective aligns with the pretraining objectives of strong semantic encoders, such as SigLIP 2 [54] and the Qwen2.5-VL [2] vision encoder. Image captioning has also been shown to provide semantic richness comparable to contrastive learning [53], 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

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.

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 [31]	7B	1510.7	62.0	54.8	65.8	35.7	33.1	55.5	17.8	31.8
Qwen-VL-Chat [1]	7B	1487.6	57.5	49.3	64.8	37.0	34.5	57.7	49.8	48.8
LLaVA-OV [23]	7B	1580.0	-	69.9	76.7	48.8	57.5	81.4	80.9	62.2
Composite UMMs										
TokenFlow-XL [42]	14B	1551.1	62.5	56.6	72.6	43.2	-	-	-	-
BLIP3-o [4]	4B	1527.7	-	60.4	73.8	46.6	-	-	-	-
Tar [15]	7B	1571.0	61.3	-	73.0	39.0	-	-	-	-
X-Omni [12]	7B	-	62.8	62.6	74.3	47.2	-	76.8	81.5	70.4
1.5B-scale Native UMMs										
Show-o [70]	1.3B	1097.2	58.0	-	51.5	27.4	-	-	-	-
Harmon [64]	1.5B	1155.0	58.9	49.8*	67.1	38.9	35.3*	57.0*	29.8*	11.2*
JanusFlow [35]	1.3B	1333.1	60.3	41.2*	**70.5**	29.3	40.6*	54.2	42.4*	53.2*
SynerGen-VL [25]	2.4B	1381.0	-	-	-	34.2	-	-	-	-
Janus-Pro [6]	1.5B	1444.0	59.3	52.6*	68.3	36.3	43.1*	64.5*	23.4	48.7
Show-o2 [71]	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 [7]	14B	1687.0	-	72.8	78.5	55.3	-	89.2	78.5	73.3
Emu3 [58]	8B	-	60.3	57.4	68.2	31.6	-	70.0	-	68.7
VILA-U [65]	7B	1401.8	60.8	-	59.0	-	-	-	-	-
MUSE-VL [72]	7B	-	-	-	69.1	39.7	49.6	69.8	-	-
Janus-Pro [6]	7B	1567.1	62.0	58.0*	72.1	41.0	48.3*	71.3*	25.8	59.0
Mogao [28]	7B	1592.0	60.9	-	74.6	44.2	-	-	-	-
Show-o2 [71]	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**

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

Datasets. For image data, we use 177M in-house image-caption pairs for Stage 1 and Stage 2 pretraining. For instruction-following in Stages 2 and 3, we include 13M conversational examples from the open-source FineVision dataset [60], together with ∼2M image-editing examples from OmniEdit [59]. In addition, our high-quality SFT corpus of ∼10M image-caption pairs combines carefully filtered real image-text pairs with a small fraction of synthetic pairs generated by existing T2I models [3, 4]. For video data, we employ ∼10M in-house video-caption pairs and ∼1.6M video instruction-following examples from LLaVA-Video-178K [81] during Stages 2 and 3 training.

Implementation details. We verify TUNA with two LLM models at different scales, *i.e.*, Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct [2]. In the pretraining stage, we optimize the representation encoder, projection layers, and diffusion head with AdamW [33] using a learning rate of 1×10^{-4}. We train on images with a base resolution of 512×512, as well as alternative resolutions that yield a similar number of visual tokens. In the second stage, we enable end-to-end training after a linear warm-up 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. Additional implementation details for each stage are provided in the appendix.

3.2. Main Results

Image and video understanding. We evaluate TUNA's multimodal understanding capabilities on nine benchmarks, including general VQA benchmarks such as MME [11], GQA [18], RealWorldQA [66] and SEED-Bench [22]; knowledge-intensive benchmarks such as MMMU [78], MMStar [5], and AI2D [20]; and text-centric benchmarks including ChartQA [36] and OCRBench [32]. As shown in Table 1, both 1.5B and 7B TUNA achieve state-of-

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

Models	Size	1-Obj.	2-Obj.	Count	Colors	Position	Col. Attr.	Overall
Generation-only Models								
SD3-M [9]	2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1 [Dev]† [3]	12B	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Composite UMMs								
MetaQuery-XL† [40]	7B	-	-	-	-	-	-	0.80
Tar [15]	7B	0.99	0.92	0.83	0.85	0.80	0.65	0.84
BLIP3-o [4]	8B	-	-	-	-	-	-	0.84
UniWorld-V1† [29]	12B	0.98	0.93	0.81	0.89	0.74	0.71	0.84
OmniGen2† [63]	7B	0.99	0.96	0.74	0.98	0.71	0.75	0.86
1.5B-scale Native UMMs								
D-DiT [27]	2B	0.97	0.80	0.54	0.76	0.32	0.50	0.65
Show-o [70]	1.5B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Janus-Pro [6]	1.5B	0.98	0.82	0.51	0.89	0.65	0.56	0.73
Show-o2 [71]	1.5B	0.99	0.86	0.55	0.86	0.46	0.63	0.73
Harmon [64]	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 [73]	7B	-	-	-	-	-	-	0.57
Transfusion [83]	7B	-	-	-	-	-	-	0.63
Emu3 [58]	8B	-	-	-	-	-	-	0.66
Show-o2 [71]	7B	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76
Janus-Pro [6]	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
BAGEL† [7]	14B	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88
Mogao [28]	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**

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. We include more video understanding results in the Appendix.

Image generation. We evaluate TUNA's image generation performance on two benchmarks: GenEval [13] and DPG-Bench [16]. Results are presented in Table 2 and Table 3. Across both 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. 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.

Image editing. We employ ImgEdit-Bench [76] as our primary evaluation suite for image editing. As shown in Table 4, TUNA achieves an overall score of 4.31, ranking highest among all UMMs. Moreover, its performance is comparable to generation-only models such as FLUX.1 Kontext [3] and Qwen-Image [62]. These results demonstrate the effectiveness of TUNA's unified architecture in handling image editing tasks, which require both precise image understanding and high-fidelity visual generation.

Video generation. We evaluate TUNA on VBench [17] for text-to-video generation, comparing it against other UMMs and generation-only models. As shown in Table 5, 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.

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

Models	Size	Global	Entity	Attribute	Relation	Other	Overall
Generation-only Models							
SD3-M [9]	2B	87.90	91.01	89.96	80.70	88.68	84.08
FLUX.1 [Dev][3]	12B	82.10	89.50	88.70	91.10	89.40	84.00
Composite UMMs							
Tar [15]	7B	83.98	88.62	88.05	93.98	84.86	84.19
BLIP3-o [4]	8B	-	-	-	-	-	81.60
UniWorld-V1[29]	12B	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2[63]	7B	88.81	88.83	90.18	89.37	90.27	83.57
1.5B-scale Native UMMs							
Harmon [64]	1.5B	-	-	-	-	-	80.93
Show-o2 [71]	1.5B	87.53	**90.38**	91.34	90.30	**91.21**	85.02
TUNA	1.5B	**88.87**	90.32	**91.71**	**91.79**	90.14	**86.03**
7B-scale Native UMMs							
Emu3-DPO [58]	8B	-	-	-	-	-	81.60
Janus-Pro [6]	7B	86.90	88.90	89.40	89.32	89.48	84.19
Mogao [28]	7B	82.37	90.03	88.26	**93.18**	85.40	84.33
BAGEL [7]	14B	88.94	90.37	**91.29**	90.82	88.67	85.07
Show-o2 [71]	7B	89.00	**91.78**	89.96	91.81	**91.64**	86.14
TUNA	7B	**90.42**	91.68	90.94	91.87	90.73	**86.76**

Table 4. Image editing results on ImgEdit benchmark. We test editing performance across various dimensions including, 'Add', 'Adjust', 'Extract', 'Replace', 'Remove', 'Background', 'Style', 'Hybrid', and 'Action'. **Bold**: best results among each section. Underline: second-best.

Models	Size	Add	Adj.	Ext.	Rep.	Rm.	Bg.	Sty.	Hyb.	Act.	Total
Generation-only Models											
FLUX.1 Kontext [3]	-	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
Qwen-Image [62]	20B	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
Native or Composite UMMs											
OmniGen [67]	3.8B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
BAGEL [7]	14B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1 [29]	12B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [63]	4B	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
GPT-Image-1 [38]	-	**4.61**	4.33	**2.90**	4.35	3.66	**4.57**	**4.93**	3.96	**4.89**	4.20
TUNA	7B	4.43	**4.48**	2.46	**4.65**	**4.55**	4.52	4.69	**4.22**	4.76	**4.31**

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 the Appendix.
3. TUNA's unified representation, initialized from three different pretrained representation encoders: SigLIP, SigLIP

Table 5. Video generation results on VBench. The complete set of full column names is included in the appendix. **Bold**: best results among each section. 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 [75]	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 [65]	7B	76.26	65.04	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	74.01
HaploOmni [68]	7B	-	-	96.40	97.60	-	96.80	65.30	-	-	-	-	-	-	-	34.60	-	-	-	78.10
Emu3 [58]	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 [71]	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. 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.

Models	ID	Data	Understanding				Generation	
			MME-p	MMMU	SEED	GQA	GenEval	DPG
Show-o2	1	Und	1351	36.1	62.1	56.8	-	-
TUNA (SigLIP 2)	2	Only	1386	37.6	62.9	57.4	-	-
Show-o2	3	Gen	-	-	-	-	76.2	82.56
TUNA (SigLIP 2)	4	Only	-	-	-	-	77.8	83.33
Show-o2	5		1339	35.5	61.7	55.9	75.9	82.32
Decoupled	6	Und	1346	37.2	61.4	56.5	78.3	83.50
TUNA (SigLIP)	7	&	1349	36.3	64.6	57.4	76.9	83.10
TUNA (SigLIP 2)	8	Gen	1361	38.1	66.5	58.2	79.4	84.20
TUNA (DINOv3)	9		1396	37.3	65.6	58.6	78.9	84.08

2, and DINOv3[3] [45].

All models are trained on a subset of our training data using a two-stage training pipeline (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 6. Detailed experimental settings and additional ablation studies are provided in the appendix.

Unified vs. decoupled visual representation. Comparing Model 6 and Model 8 in Table 6, we observe that our unified representation consistently outperforms the decoupled setting across all understanding and generation benchmarks. This result demonstrates the challenge of jointly processing multiple visual representations within a single model, and highlights the necessity and effectiveness of adopting a unified representation for building robust UMMs.

Selection of representation encoders. We find that TUNA's unified representation generally benefits from stronger representation encoders. As shown in Table 6, comparing Models 7, 8, and 9, both SigLIP 2 (400M parameters) and DINOv3 (800M) outperform SigLIP (400M) on all benchmarks. We 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. As shown in Table 6, we further conduct two additional experiments for TUNA and Show-o2, where each model is trained exclusively on either understanding (Models 1 and 2) or generation data (Models 3 and 4). The results demonstrate that TUNA benefits from joint training on both data types: Model 8 surpasses Model 2 on understanding benchmarks and Model 4 on generation benchmarks, highlighting the mutual reinforcement between understanding and generation within our unified framework.

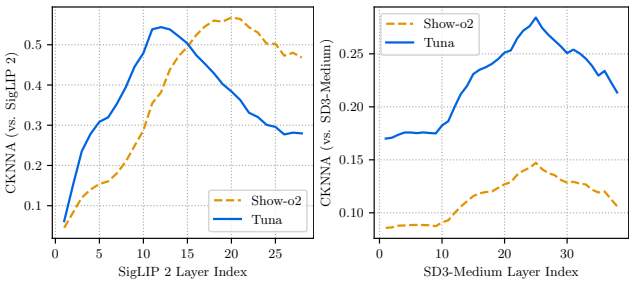
Comparison with Show-o2. A closely related work to ours is Show-o2 [71], which uses a dual-path late-fusion mechanism, merging features from separate VAE and semantic encoders via a fusion layer to build unified representations. In contrast, TUNA extracts unified representations directly from VAE latents using a semantic encoder, achieving deep feature fusion across all layers in the semantic encoder. Comparing Model 5 with Models 7 and 8 in Table 6, our unified representation consistently outperforms Show-o2 on all benchmarks, regardless of the semantic encoder. Model 1 vs. 2 and Model 3 vs. 4 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

Our ablation results in Section 3.3 show that TUNA's unified representation outperforms the dual-path late-fusion strategy of Show-o2 across both understanding and generation benchmarks. To better understand why our unified representation yields superior performance, we perform a representation alignment analysis using CKNNA scores [19] with respect to two reference models: (1) a strong semantic encoder SigLIP 2 [54], and (2) a strong generation model SD3-Medium [9]. Concretely, we extract unified visual representations from TUNA and Show-o2 based on 1,024 images from the Wikipedia Captions dataset [47] and compute their CKNNA scores relative to features from all intermediate layers of the two reference models. The results are presented in Figure 4a and Figure 4b.

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

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



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

Figure 4. 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.

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 and generation 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 unified design learns a single shared representation optimized jointly for both tasks, effectively eliminating this imbalance and yielding superior performance in both understanding and generation.

4. Related Work

4.1. 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 with a composite design (composite UMMs) such as MetaQuery [40], BLIP-3o [4], and UniWorld-V1 [29] 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 [6, 35, 61] and UniFluid [10] adopt decoupled visual representations for understanding and generation. BAGEL [7], Mogao [28], and OneCAT [24] further use MoE-style architectures to route inputs separately, mitigating conflicts between distinct representations from the decoupled vision encoders. Alternatively, models like Transfusion [83], Harmon [64], and the Show-o series [70, 71] 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.

4.2. Representation in Unified Models

Recent research has primarily focused on developing unified visual tokenizers that support both understanding and generation tasks. For instance, TokenFlow [42] and MUSE-VL [72] adopt late-fusion strategies to merge features from separate understanding and generation encoders into quantized codebooks. DualToken [46], UniTok [34] and TokLIP [30] 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 [79] and UniLIP [50] 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 simply learns a unified representation end-to-end under joint understanding and generation objectives, and is validated at larger scale across more tasks. Moreover, while UniLIP adopts a composite design where the unified features only serve as conditions for a separate pretrained generative model (SANA [69]), 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.

References

[1] 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. 5

[2] 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. 3, 4, 5

[3] 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. 3, 5, 6

[4] 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*, 2025. 5, 6, 8

[5] 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, 2024. 2, 5

[6] 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*, 2025. 5, 6, 8

[7] 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. 2, 4, 5, 6, 8

[8] 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. 2

[9] 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. 3, 6, 7

[10] 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. 8

[11] 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, 2025. 5

[12] 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. 5

[13] 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. 2, 6

[14] 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. 3

[15] 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. 5, 6

[16] 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. 6

[17] 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. 6

[18] 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. 5

[19] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 7

[20] 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. 5

[21] 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, 2023. 3

[22] 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*, 2023. 5

[23] 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*, 2024. 5

[24] 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*, 2025. 8

[25] 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, 2025. 5

[26] 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, 2024. 2

[27] 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, 2025. 6

[28] 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. 2, 5, 6, 8

[29] 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. 6, 8

[30] 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, 2025. 3, 8

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5

[32] 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, 2024. 5

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[34] 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*, 2025. 3, 8

[35] 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, 2025. 5, 8

[36] 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. 5

[37] 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. 3

[38] OpenAI. gpt-image-1. Model documentation. 6

[39] 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. 3

[40] 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. 6, 8

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4

[42] 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. 5, 8

[43] 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. 3

[44] 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. 4

[45] 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. 7

[46] 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. 8

[47] 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. 7

[48] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[49] 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. 3

[50] 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. 8

[51] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[52] Wan-Video team. Wan: Open and advanced large-scale video generative models (wan2.2), 2025. GitHub repository, Apache-2.0 License. 4

[53] 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. 4

[54] 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. 4, 7

[55] 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. 2, 3

[56] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024. 2

[57] 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*, 2025. 3

[58] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 5, 6, 7

[59] 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. 5

[60] 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. 5

[61] 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, 2025. 8

[62] 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*, 2025. 6

[63] 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*, 2025. 6

[64] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025. 2, 5, 6, 8

[65] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 5, 7

[66] xAI. Grok-1.5 vision preview. Company news post. 5

[67] 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. 6

[68] Yicheng Xiao, Lin Song, Rui Yang, Cheng Cheng, Zunnan Xu, Zhaoyang Zhang, Yixiao Ge, Xiu Li, and Ying Shan. Haploomni: Unified single transformer for multimodal video understanding and generation. *arXiv preprint arXiv:2506.02975*, 2025. 7

[69] 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*, 2024. 8

[70] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 4, 5, 6, 8

[71] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 2, 4, 5, 6, 7, 8

[72] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. *arXiv preprint arXiv:2411.17762*, 2024. 5, 8

[73] 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, 2025. 6

[74] 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. 3

[75] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 7

[76] 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. 6

[77] 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. 2, 3

[78] 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. 5

[79] 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. 8

[80] 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. 2

[81] 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*, 2024. 5

[82] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 3

[83] 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. 2, 6, 8

[84] 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. 3

[85] 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. 3