TUNA-2: Pixel Embeddings Beat Vision Encoders for Multimodal Understanding and Generation

Zhiheng Liu[♡◇]*, **Weiming Ren**[♡♠]*, **Xiaoke Huang**[♡], **Shoufa Chen**[♡], **Tianhong Li**[♡], **Mengzhao Chen**[◇], **Yatai Ji**[◇], **Sen He**[♡], **Jonas Schult**[♡], **Belinda Zeng**[♡], **Tao Xiang**[♡], **Wenhu Chen**[♠], **Ping Luo**[◇], **Luke Zettlemoyer**[♡], **Yuren Cong**[♡]
Meta AI[♡], The University of Hong Kong[◇], University of Waterloo[♠]
Project Page: https://tuna-ai.org/tuna-2/

Abstract

Unified multimodal models typically rely on pretrained vision encoders and use separate visual representations for understanding and generation, creating misalignment between the two tasks and preventing fully end-to-end optimization from raw pixels. We introduce TUNA-2, a native unified multimodal model that performs visual understanding and generation directly based on pixel embeddings. TUNA-2 drastically simplifies the model architecture by employing simple patch embedding layers to encode visual input, completely discarding the modular vision encoder designs such as the VAE or the representation encoder. Experiments show that TUNA-2 achieves state-of-the-art performance in multimodal benchmarks, demonstrating that unified pixel-space modelling can fully compete with latent-space approaches for high-quality image generation. Moreover, while the encoder-based variant converges faster in early pretraining, TUNA-2's encoder-free design achieves stronger multimodal understanding at scale, particularly on tasks requiring fine-grained visual perception. These results show that pretrained vision encoders are not necessary for multimodal modelling, and end-to-end pixel-space learning offers a scalable path toward stronger visual representations for both generation and perception.

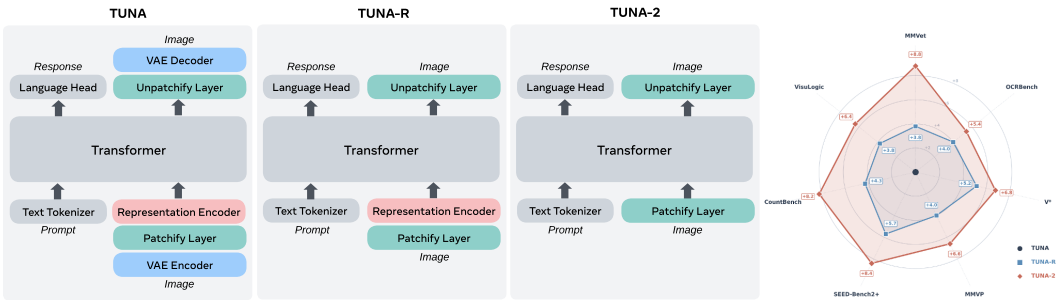

Figure 1: Evolution of TUNA-2 architecture and multimodal performance comparison. We simplify existing UMMs (*e.g.*, TUNA [60]) by progressively stripping away their visual encoding components. By removing the VAE, we first derive TUNA-R, a pixel-space UMM that relies solely on a representation encoder. TUNA-2 further streamlines the design by bypassing the representation encoder entirely, utilizing direct patch embedding layers for raw image inputs. TUNA-2 using pixel embeddings outperforms both TUNA-R and TUNA across a diverse suite of multimodal benchmarks.

*Joint first authors, listed alphabetically by last name.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Visual understanding and generation are two core capabilities in multimodal AI. Recent work has increasingly focused on native unified multimodal models (UMMs) [128, 23, 60], which aim to integrate both capabilities within a single framework. A central challenge in building such models is encoding input images into visual representations that effectively support both understanding and generation. Early approaches [23, 19] adopted decoupled representations, using representation encoders such as CLIP [72] for understanding and reconstruction-oriented encoders such as VQ-VAE [27] for generation. To address the representation mismatch introduced by this design, more recent UMMs [108, 60] have moved toward modelling both tasks using unified visual representations through a shared vision encoder.

Despite the significant progress, both decoupled and unified visual representation designs still rely heavily on pretrained vision encoders [86, 85] for visual feature extraction. In parallel, recent research on multimodal understanding and generation has begun to move away from encoder-based modular designs toward simpler monolithic, encoder-free architectures. In multimodal understanding, newer native vision-language models [26] remove the pretrained representation encoder and instead align images and natural language within a unified, end-to-end architecture. In visual generation, pixel-space diffusion models [40, 18, 50] have shown increasing flexibility, stronger scalability, and state-of-the-art performance on a wide range of tasks, suggesting that pretrained VAE encoders may no longer be essential even for high-fidelity image synthesis.

Motivated by these observations, we ask a natural but largely unexplored question: can we move beyond pretrained vision encoders altogether, and build unified multimodal models through end-to-end native modelling directly from raw pixels?

We answer this question affirmatively by introducing TUNA-2, a native unified multimodal model that attempts to progressively simplify the encoder modules, and ultimately remove vision encoders completely. We first introduce TUNA-R, which eliminates the VAE model while keeping a representation encoder in the model architecture. TUNA-R performs multimodal understanding similar to standard encoder-based LMMs, and supports visual generation through pixel-space flow matching with an x-prediction objective. We then propose TUNA-2, which further simplifies the architecture by removing the encoder entirely and using only a single transformer decoder to process image and video tokens. As a result, TUNA-2 enables end-to-end native unified modelling directly from raw pixels, without relying on any pretrained encoder modules.

Since learning unified representations directly in high-dimensional pixel space is substantially more challenging than learning them in a compact latent space, we further introduce a masking-based visual feature learning scheme to stabilize training and encourage the learning of more robust pixel-space representations. Together, these designs enable TUNA-2 to achieve state-of-the-art performance across a diverse set of multimodal understanding and generation benchmarks. More importantly, our controlled comparison reveals a clear design insight: after sufficient visual pretraining, the encoder-free TUNA-2 becomes competitive with the encoder-based TUNA-R on visual generation, while consistently outperforming it on multimodal understanding, especially on benchmarks that require fine-grained visual perception. These findings suggest that removing pretrained vision encoders can be advantageous for learning stronger fine-grained visual representations in end-to-end pretraining. As shown in Figures 1 and 2, this leads to highly competitive performances in both multimodal understanding and generation.

Our main contributions are summarized as follows:

1. We propose TUNA-2, a native unified multimodal model that supports multimodal understanding and generation with encoder-free designs, achieving state-of-the-art performance across a wide range of understanding and generation benchmarks.
2. We conduct controlled comparisons between TUNA-2 and an encoder-based pixel-space UMM variant TUNA-R, and show that after sufficient multimodal pretraining, TUNA-2 and its encoder-free design are competitive on generation and advantageous for understanding, especially on fine-grained, perception-intensive tasks.
3. We conduct comprehensive ablations and analyses on pixel-space UMMs to study their training dynamics and model behaviours, offering useful insights for the development of future native unified multimodal models.

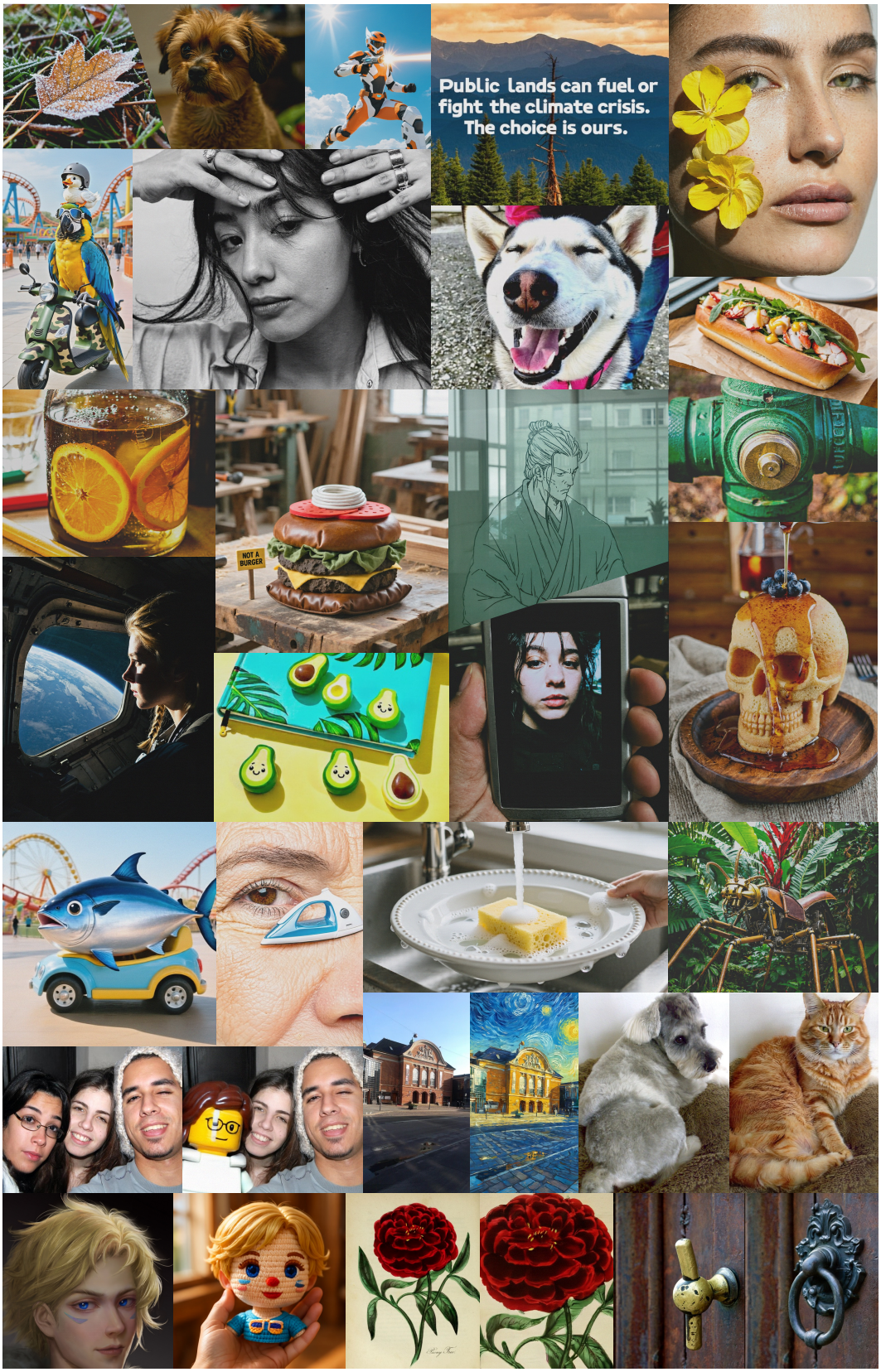

Figure 2: While being completely encoder-free, TUNA-2 is capable of performing high-fidelity text-to-image generation and image editing.

2 Method

In this section, we present TUNA-2, a native unified multimodal model that performs visual understanding and generation both in pixel space. We start by detailing our approach to progressively remove vision encoder components to derive TUNA-2 in Section 2.1. We then describe our masked feature learning scheme in Section 2.2 and our model training pipeline in Section 2.3.

2.1 Towards Encoder-Free Unified Models

As shown in Figure 1, existing UMMs with unified visual representations, such as TUNA [60], typically consist of a vision encoder and an LLM decoder for joint vision-language modeling, followed by modality-specific heads, including a language modelling head for autoregressive text generation and a flow matching head for image generation. In this work, we propose TUNA-2 as an encoder-free UMM formulation by progressively simplifying the vision encoder components in existing architectures. Our design process for this architectural simplification is as follows:

Representation encoder-based architecture. First, we attempt to remove the VAE model and only employ a pretrained representation encoder in the vision encoder. As shown in Figure 1, this resonates a standard paradigm for vision-language modelling: the representation encoder first encodes input images into visual tokens, which are then combined with the text tokens in the LLM decoder for joint vision-language modelling. Originally proposed in LLaVA [57], this paradigm has been verified and scaled up by later works such as Qwen3-VL [9] and InternVL3.5 [90], and remains the most popular framework for multimodal understanding. We refer to this intermediate design as **TUNA-R**. Although our ultimate goal is to move beyond encoder-based architectures, we view TUNA-R as an important intermediate step that enables a controlled comparison with TUNA-2.

Encoder-free (non-encoder) architecture. Second, we consider a further simplified architecture that removes the representation encoder entirely, which becomes our main design for **TUNA-2**. As shown in Figure 1, this design replaces pretrained vision encoders with simple patch embedding layers that convert images into visual tokens, which are then processed jointly with text tokens by the LLM decoder. Similar encoder-free designs have recently been explored in models such as Mono-InternVL [62] and NEO [26]. By removing the pretrained representation encoder, this design avoids its built-in inductive biases, such as fixed input resolutions and limited access to fine-grained low-level visual details. It also simplifies the model architecture into a single unified transformer. In Section 3, we present a series of in-depth analyses comparing TUNA-2 with TUNA-R, and demonstrate the effectiveness and scalability of TUNA-2.

Pixel-space image generation. Our VAE-free design allows us to directly perform multimodal understanding and text generation using the LLM decoder and the language modelling head. However, discarding the VAE also means that we can no longer adopt the designs from existing UMMs and generation-only models that follow the latent diffusion architecture. To effectively perform pixel-space image generation, we adopt the x-prediction and v-loss paradigm from JiT [50] for pixel-space flow matching. Specifically, given the source image x_1, the sampled noise $x_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the timestamp t, we employ rectified flow [58, 56] and its linear schedule to construct a noisy sample in pixel space:

$$x_t = tx_1 + (1-t)x_0, t \in [0, 1]. \tag{1}$$

TUNA-2 is then formulated to directly predict the clean image from the noisy image in pixel space:

$$x_\theta = \pi_\theta(x_t, c, t), \tag{2}$$

where π_θ is our unified model (vision-language backbone and flow matching head) and c is the conditioning signals (text for text-to-image generation and text+image for image editing). As suggested in JiT, while our model directly predicts x_θ, we still transform it into the velocity term v_θ and regress v_θ as our learning objective:

$$v_\theta = \frac{x_\theta - x_t}{1-t}, \tag{3}$$

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,c,x_1,x_0}||v_\theta - v||_2^2, \tag{4}$$

where v is the ground truth velocity defined by $v = x_1 - x_0$. During inference, we employ the Euler solver and predict the denoised image at t' from the noisier image at $t < t'$ based on the velocity term v_θ, such that $x_{t'} = x_t + (t' - t)v_\theta$, where v_θ is transformed from our model prediction x_θ, based on Equation 3.

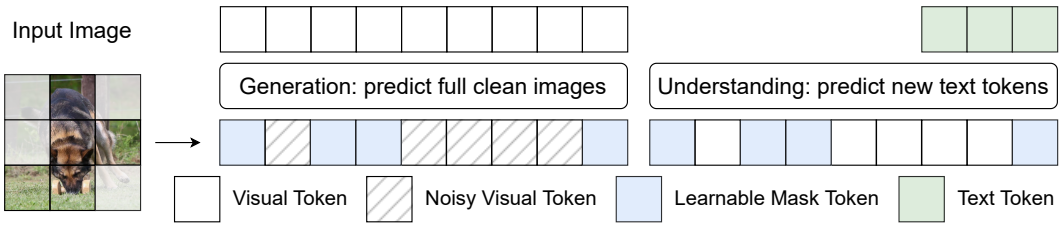

Figure 3: Illustration of our proposed masking-based feature learning scheme. During training, we use the learnable mask token to regularize multimodal understanding and perform masked prediction for visual generation.

2.2 Learning Robust Visual Representations via Masking

While removing the VAE simplifies our model architecture and enables fully end-to-end unified multimodal training, it also shifts visual modelling from a compact latent space to the much higher-dimensional pixel space. As a result, learning a unified visual representation becomes more challenging: the increased redundancy in pixel-space inputs makes it easier for the model to rely on superficial shortcuts, rather than learning visual cues that are genuinely informative for both understanding and generation. To learn more robust visual representations in pixel space, we introduce a masking-based visual feature learning scheme. As shown in Figure 3, during training, we (optionally) randomly select a subset of image patches according to a masking ratio and replace the masked visual tokens with a learnable mask token before feeding them into the LLM decoder. The same masking operation is applied to both generation and understanding examples, but plays different roles in the two settings:

1. For generation examples, we let the model predict the clean image patches in *both the masked and the unmasked regions*, such that (1) we create a harder denoising problem for the model to predict clean images from partially observed noisy images; and (2) it encourages the learnable mask token to absorb useful information for image reconstruction conditioned on the visible context.
2. For understanding examples, our model predicts the ground truth text response based on the *masked visual input*. In this case, masking serves as a regularization mechanism that forces the model to perform multimodal reasoning under partial visual observation, leading to more robust visual representations.

Our masking-based feature learning scheme resembles masked modelling methods in visual understanding and generation, such as MAE [37] and SigLIP 2 [85] for semantic learning and MaskGIT [15] and DeTok [113] for visual generation. Empirically, we find that applying masking leads to enhanced model performance during pretraining stages.

2.3 Training Pipeline

Our encoder-free design enables fully end-to-end training of TUNA-2, without requiring separate stages to train connector layers, which is a common design in encoder-based modular approaches. As described below, our training pipeline consists of two stages, both of which are carried out in a fully end-to-end manner:

Stage 1: full model pretraining. In the first stage, we aim to establish a strong initialization for the flow matching head, and adapt pixel-space visual inputs for unified multimodal understanding and generation. To achieve this, we train the full model jointly on two tasks: image captioning and text-to-image generation.

Stage 2: supervised finetuning (SFT). Next, we perform supervised fine-tuning (SFT) of the full model with a lower learning rate. We use datasets for image editing, image instruction-following, and high-quality image generation. This step refines TUNA-2's abilities, boosting performance and generalization across various multimodal tasks.

For TUNA-R, which includes a connector layer between the representation encoder and the LLM decoder, we add an extra alignment stage before Stage 1. In this stage, we train only the connector layer for a short period using image captioning and text-to-image generation data. As noted above, TUNA-2 does not require this additional stage because of its encoder-free design.

Table 1: Comparisons between TUNA-2 and baseline models on multimodal understanding benchmarks. Results with model size greater than 13B are grayed. **Bold**: best results among all UMMs. Underline: second-best among all UMMs.

Models	Size	General Benchmarks									Pixel-centric Benchmarks		
		GQA	RealWorldQA	MMVet	MMMU	MMVP	SEED-Bench2+	AI2D	ChartQA	OCRBench	V*	CountBench	VisuLogic
Understanding-only Models (LMMs)													
LLaVA-1.5 [57]	7B	62.0	54.8	32.9	35.7	-	-	55.5	17.8	31.8	-	-	-
Qwen-VL-Chat [8]	7B	57.5	49.3	47.3	37.0	-	-	57.7	49.8	48.8	-	-	-
LLaVA-OV [47]	7B	-	69.9	51.9	48.8	77.3	62.2	81.4	80.9	62.2	72.7	76.2	24.8
Qwen2.5-VL [47]	7B	60.7	69.9	61.7	58.6	78.0	70.5	82.7	83.0	83.7	71.2	74.1	20.0
Composite UMMs													
TokenFlow-XL [70]	14B	62.5	56.6	-	43.2	-	-	-	-	-	-	-	-
BLIP3-o [16]	4B	-	60.4	-	46.6	-	-	-	-	-	-	-	-
Tar [35]	7B	61.3	-	-	39.0	74.3	46.2	-	-	-	41.4	64.2	24.3
X-Omni [32]	7B	62.8	62.6	-	47.2	-	-	76.8	81.5	70.4	-	-	-
Native UMMs													
BAGEL [23]	14B	66.4	72.8	67.2	55.3	85.0	71.9	89.2	78.5	73.3	70.2	82.5	41.7
Ming-UniVision [42]	16B	59.4	59.1	64.2	40.3	71.0	56.8	82.8	76.7	72.4	48.2	76.8	26.7
PixelUMM [?]	16B	61.9	71.6	-	41.7	80.3	64.5	80.1	83.0	78.0	77.0	94.3	-
Harmon [102]	1.5B	58.9	49.8	-	38.9	61.7	41.6	57.0	29.8	11.2	41.9	67.0	25.1
JanusFlow [64]	1.3B	60.3	41.2	36.2	29.3	67.7	39.8	54.2	42.4	53.2	42.9	78.6	22.0
Emu3 [92]	8B	60.3	57.4	23.5	31.6	71.0	44.6	70.0	69.4	68.7	53.4	65.2	24.7
VILA-U [103]	7B	60.8	46.8	26.3	31.2	62.7	31.9	49.0	11.4	23.3	38.7	55.2	25.4
Janus-Pro [19]	7B	62.0	58.0	41.1	41.0	73.3	56.3	71.3	25.8	59.0	47.6	53.2	23.8
Show-o2 [108]	7B	63.1	64.7	39.6	48.9	76.7	61.3	78.6	52.3	32.4	44.5	63.5	26.9
OneCat [49]	9B	63.1	65.2	**52.2**	41.9	71.3	**61.6**	77.8	81.2	79.0	**63.4**	34.2	24.9
TUNA [60]	7B	63.9	66.1	42.9	49.8	70.7	52.7	79.3	**85.8**	74.3	52.4	73.5	22.4
TUNA-R	7B	63.5	**67.9**	46.7	**51.1**	74.7	58.4	79.4	85.6	78.3	57.6	77.8	26.2
TUNA-2	7B	**65.0**	67.7	51.7	50.7	**77.3**	61.1	**79.6**	85.6	**79.7**	59.2	**81.7**	**28.8**

3 Experiments

3.1 Experiment Setup

We employ Qwen2.5-7B-Instruct [71] as the LLM decoder and use a patch embedding size of 16 for TUNA-2. For Stage 1 pretraining, we use 550M in-house image-text pairs, consisting of 70% image captioning data for multimodal understanding and 30% text-to-image generation data. In addition, we include text-only data from Nemotron [13], which accounts for 20% of the total pretraining data. The full model is trained end-to-end for 300k steps on 64 nodes with the AdamW optimizer [61] and a learning rate of 1×10^{-4}. We employ the proposed masking-based feature learning strategy during the final 40% of pretraining, applying masking in 50% of training examples and randomly sampling the masking ratio from 0% to 50%. For Stage 2 supervised finetuning, we use a curated SFT corpus covering image instruction-following, image editing, and high-quality image generation. Specifically, for image instruction-following, we include 13M conversational examples from the open-source FineVision [97] dataset. For image editing, we use approximately 2M examples from OmniEdit [94]. This stage is trained for 50k steps with AdamW and a learning rate of 2×10^{-5}. For all training stages, we pad the input sequence length to 16K tokens per GPU.

For TUNA-R, we use the same Qwen2.5-7B-Instruct as the LLM decoder. We follow TUNA and adopt SigLIP 2 So400M [85] as the representation encoder. For the connector-alignment stage in TUNA-R, we train the model for 3k steps with AdamW and a learning rate of 5×10^{-4}.

3.2 Main Results

Image understanding. We employ a comprehensive evaluation suite consisting of nine VQA benchmarks, including GQA [44], RealWorldQA [104], MMVet [119], MMMU [121], MMVP [83], SEED-Bench2+ [48], AI2D [45], ChartQA [65] and OCRBench [59], to evaluate the image understanding capabilities for TUNA-2. As shown in Table 1, after removing the VAE model, both TUNA-R and TUNA-2 outperform TUNA and achieve state-of-the-art results among all 7B-scale native UMMs, demonstrating the effectiveness of our pixel-space unified representations. Notably, TUNA-2 outperforms TUNA-R even after replacing the representation encoder with the simple patchify layer. This may indicate that large-scale joint training of a unified, monolithic architecture achieves better understanding performance than relying on the inductive bias in the pretrained representation encoders in the modular setting. We show more analysis comparing TUNA-R and TUNA-2 in Section 3.4.

To further understand the benefit of our pixel-space unified models, we include several pixel-centric benchmarks that focus heavily on visual reasoning over fine-grained visual details (e.g. recognizing

Table 2: Image generation results on GenEval and DPG-Bench. "Col. Attr." means "Color Attribute". † refers to methods using LLM rewriters in GenEval. **Bold**: best results among native UMMs. Underline: second-best.

Models	Size	GenEval							DPG-Bench					
		1-Obj.	2-Obj.	Count	Colors	Position	Col. Attr.	Overall	Global	Entity	Attribute	Relation	Other	Overall
Generation-only Models														
SD3-M [28]	2B	0.99	0.94	0.72	0.89	0.33	0.60	0.74	87.90	91.01	89.96	80.70	88.68	84.08
FLUX.1 [dev]† [11]	12B	0.98	0.93	0.75	0.93	0.68	0.65	0.82	82.10	89.50	88.70	91.10	89.40	84.00
LongCat-Image [81]	6B	0.99	0.98	0.86	0.86	0.75	0.73	0.87	89.10	92.54	92.00	93.28	87.50	86.80
Qwen-Image [98]	20B	0.99	0.92	0.89	0.88	0.76	0.77	0.87	91.32	91.56	92.02	94.31	92.73	88.32
Seedream 3.0 [31]	-	0.99	0.96	0.91	0.93	0.47	0.80	0.84	94.31	92.65	91.36	92.78	88.24	88.27
Z-Image-Turbo [14]	6B	1.00	0.95	0.77	0.89	0.65	0.68	0.82	91.29	89.59	90.14	92.16	88.68	84.86
Composite UMMs														
Tar [35]	7B	0.99	0.92	0.83	0.85	0.80	0.65	0.84	83.98	88.62	88.05	93.98	84.86	84.19
BLIP3-o [16]	8B	-	-	-	-	-	-	0.84	-	-	-	-	-	81.60
UniWorld-V1† [54]	12B	0.98	0.93	0.81	0.89	0.74	0.71	0.84	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2† [99]	7B	0.99	0.96	0.74	0.98	0.71	0.75	0.86	88.81	88.83	90.18	89.37	90.27	83.57
Native UMMs														
MUSE-VL [109]	7B	-	-	-	-	-	-	0.57	-	-	-	-	-	-
Transfusion [128]	7B	-	-	-	-	-	-	0.63	-	-	-	-	-	-
Emu3 [92]	8B	-	-	-	-	-	-	0.66	-	-	-	-	-	81.60
Show-o2 [108]	7B	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76	89.00	91.78	89.96	91.81	91.64	86.14
Janus-Pro [19]	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80	86.90	88.90	89.40	89.32	89.48	84.19
HBridge† [91]	7B	**1.00**	0.96	0.80	0.94	0.77	0.78	0.87	**91.78**	**91.82**	90.23	90.06	88.42	85.23
Ming-UniVision [42]	16B	**1.00**	0.93	0.59	0.93	**0.92**	0.70	0.85	-	-	-	-	-	82.12
BAGEL† [23]	14B	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88	88.94	90.37	91.29	90.82	88.67	85.07
Mogao [53]	7B	**1.00**	**0.97**	0.83	0.93	0.84	0.80	0.89	82.37	90.03	88.26	93.18	85.40	84.33
PixelUMM [**?**]	16B	0.99	0.95	0.65	0.84	0.55	0.64	0.77	88.87	88.92	90.00	92.53	**92.62**	85.74
TUNA [60]	7B	**1.00**	**0.97**	0.81	0.91	0.88	**0.83**	**0.90**	90.42	91.68	90.94	91.87	90.73	**86.76**
TUNA-R	7B	**1.00**	0.95	0.82	0.89	0.86	0.79	0.88	86.00	91.80	91.03	**93.48**	84.89	86.35
TUNA-2	7B	0.99	0.96	0.80	0.91	0.84	0.76	0.87	89.50	91.40	**92.07**	91.91	88.81	86.54

very small objects in high-resolution images). These benchmarks include V* [100], CountBench [68] and VisuLogic [111]. As shown in Table 1, both TUNA-R and TUNA-2 outperform latent-space UMMs (e.g. Show-o2, TUNA, etc.) across all benchmarks, indicating the necessity of pixel-space visual representations when reasoning over fine-grained visual details.

Image generation. We evaluate the image generation performance of TUNA-2 on GenEval [33] and DPG-Bench [41]. As shown in Table 2, both TUNA-R and TUNA-2 achieve state-of-the-art results on these benchmarks and perform competitively with contemporary approaches such as BAGEL, Mogao, and TUNA. TUNA-R performs slightly better than TUNA-2, suggesting that the semantic prior introduced by the representation encoder helps our method learn a stronger image generation model. Overall, these results show that even without a VAE and while performing image generation entirely in pixel space, TUNA-2 remains competitive with recent state-of-the-art unified models, highlighting the effectiveness and scalability of our pixel-space generation design.

Since existing image generation benchmarks such as GenEval mainly evaluate text-image alignment and the model's world knowledge, we further adopt an LLM-judge-based evaluation to assess the quality and diversity of images generated by TUNA-2, TUNA-R, and TUNA. Specifically, we sample 1.5K text prompts and ask each model to generate four images per prompt. We then use GPT-5.4 [67] and Claude Opus 4.7 [7] as judges to select the best model among the three for each prompt based on the quality and diversity of the generated images, where quality mainly refers to realism and fine-grained detail and texture fidelity, while diversity measures the extent to which the four generated images exhibit distinct visual variations under the same prompt. The results are reported in Table 3. We find that under both LLM judges, TUNA-2 achieves competitive image generation quality, performing comparably to TUNA-R and better than TUNA, while being significantly preferred in terms of diversity. These results demonstrate that our encoder-free design is highly effective and enables TUNA-2 to generate both high-quality and diverse images.

Image editing. We further evaluate the image editing capability of TUNA-2 on ImgEdit [116]. As shown in Table 4, TUNA-2 achieves strong editing performance among unified models, outperforming earlier baselines such as OmniGen [105], BAGEL [23], UniWorld [54], and OmniGen2 [99]. While

Table 3: Evaluation results under GPT-5.4 and Claude Opus 4.7. Quality measures image realism and fine-grained detail richness, while Diversity measures variation among images generated from the same prompt. **Bold**: best results. Underline: second-best.

	GPT-5.4		Claude Opus 4.7	
Models	Quality	Diversity	Quality	Diversity
TUNA	22.3%	20.6%	28.1%	28.2%
TUNA-R	**35.7%**	30.9%	**37.2%**	29.9%
TUNA-2	32.1%	**48.4%**	34.8%	**41.9%**

Table 4: Image editing results on ImgEdit. **Bold** denotes the best result among unified models and underlining denotes the second-best.

Models	Add	Adj.	Ext.	Rep.	Rm.	Bg.	Sty.	Hyb.	Act.	Total
Generation-only Models										
FLUX.1	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
Qwen-Image	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
Unified Models										
OmniGen	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
BAGEL	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
GPT-Image	**4.61**	4.33	**2.90**	4.35	3.66	**4.57**	**4.93**	3.96	**4.89**	4.20
TUNA	4.43	**4.48**	2.46	**4.65**	**4.55**	4.52	4.69	**4.22**	4.76	**4.31**
TUNA-R	4.46	4.27	2.38	4.61	4.48	4.44	4.54	4.06	4.43	4.18
TUNA-2	4.34	4.13	2.22	4.53	4.42	4.36	4.58	3.91	4.28	4.09

being slightly behind TUNA and TUNA-R, TUNA-2 remains competitive with strong generation-only and unified editing systems, despite performing editing directly in pixel space without relying on vision encoders. These results suggest that our pixel-space unified modelling framework can effectively support instruction-guided image editing, while the small gap between TUNA-2 and encoder-based variants indicates that pretrained visual priors may still provide benefits for fine-grained editing fidelity.

3.3 Ablation: Masking-based Feature Learning

Table 5: Ablation study results for TUNA-R and TUNA-2, with or without our proposed masking-based feature learning scheme during pretraining.

Models	OCRBench	MMVP	CountBench	GenEval
TUNA	56.9	54.0	55.6	57.2
TUNA-R w/o Masking	58.3	56.7	57.2	55.7
TUNA-R w/ Masking	59.2	58.0	58.2	56.0
TUNA-2 w/o Masking	55.4	52.3	53.4	47.6
TUNA-2 w/ Masking	56.8	55.7	57.6	48.2

To verify the effectiveness of our proposed masking-based feature learning strategy, we conduct controlled experiments on both TUNA-R and TUNA-2 using the smaller Qwen-2.5-Instruct-1.5B backbone. Since we expect this strategy to serve as a representation enhancement strategy after the model has acquired basic multimodal knowledge, we do not apply this objective from the beginning of pretraining. Instead, we first train all models for 50k steps; We then split each model variant into two groups: one continues standard pretraining, while the other continues pretraining with the masking-based feature learning objective applied with a probability of 50%. Both groups are further trained for another 50k steps. As shown in Table 5, this strategy consistently improves performance on both understanding and generation benchmarks for both model variants. We further observe that TUNA-2 benefits more from masked training than TUNA-R. We hypothesize that this difference is related to the SigLIP 2 representation encoder used in TUNA-R, since SigLIP 2 itself is pretrained with a similar masked prediction objective. In addition, the results suggest that compressing images into the VAE latent space before visual encoding (TUNA) can introduce certain information loss for visual understanding, compared with directly encoding pixel-level inputs through the vision encoder (TUNA-R). Based on these findings, we apply the masking-based feature learning strategy during the final 40% of pretraining, encouraging more robust visual representation learning for both multimodal understanding and generation.

3.4 Analysis: TUNA-R vs. TUNA-2

Our experimental results in Section 3.2 reveal two interesting findings: First, without vision encoders, TUNA-2 outperforms TUNA-R on most image understanding benchmarks. Second, TUNA-R achieves similar or slightly stronger performance than TUNA-2 on generation tasks. To better understand these trends, we plot the performance curves of both models on understanding and generation benchmarks

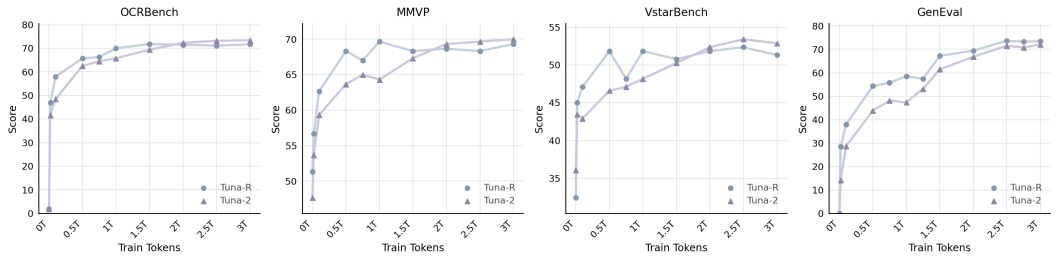

Figure 4: Model accuracy curves with respect to training dataset size (measured by the number of tokens consumed by the model) for TUNA-R and TUNA-2.

as the training data scale increases. As shown in Figure 4, on the three understanding benchmarks (OCRBench [59], MMVP [83] and V* [100]), TUNA-R actually outperforms TUNA-2 during the early stage of the training. We believe this advantage comes from the pretrained representation encoder in TUNA-R, whose rich semantic priors help the model acquire multimodal understanding capabilities more quickly at the beginning of training. Nevertheless, TUNA-2 later catches up and eventually surpasses TUNA-R. This suggests that the monolithic, encoder-free design of TUNA-2 may be better suited to benefit from large-scale unified multimodal pretraining, enabling it to develop stronger multimodal understanding capabilities.

For generation evaluation on GenEval [33], we observe that TUNA-R consistently outperforms TUNA-2 throughout the entire training process. This suggests that the semantic priors provided by the representation encoder play an important role in improving generation performance, which is consistent with prior findings from REPA [118] and TUNA [60]. However, this trend gradually weakens as the training data scale increases. As shown in the figure, the two model variants achieve nearly identical performance after SFT. Overall, compared with latent-space generation, both TUNA-R and TUNA-2's pixel-space generation paradigm achieves competitive performances. We include more analysis on the cross-modal alignment capabilities for both TUNA-R and TUNA-2 in Appendix D.

4 Related Works

In this section, we provide a brief review of unified multimodal models. We present an additional literature review on encoder-free multimodal understanding and generation models in Appendix A.

Unified models (UMMs) [110, 117, 82, 84, 124, 114, 77, 76, 106, 66, 95, 93, 96, 22, 38, 87, 36, 81, 1, 101, 29] seek to integrate multimodal understanding and generation within a single framework. A prevalent approach for UMMs combines autoregressive (AR) language models for text generation [34, 8] with diffusion or flow-matching models for visual generation [39, 74]. While AR models excel at understanding, their integration with high-fidelity generation typically relies on decoupled encoders [19, 53, 23], leading to representation mismatches and inefficiencies. To address this problem, recent works such as UniTok [63], TokLip [55], UniLip [79], UniFlow [122], UAE [30] and OpenVision 3 [125] pretrain unified vision tokenizers for both semantic understanding and visual reconstruction. Meanwhile, native UMMs like the Show-o series [107, 108], TUNA [60], Ming-UniVision [43], and Transfusion-RAE [84] build unified representations using pretrained VAE [86] and representation encoders [127]. These approaches enhance model performances, despite still relying on VAE latents.

5 Conclusion

We introduced TUNA-2, a family of native unified multimodal models that perform multimodal understanding and visual generation directly in pixel space, without relying on VAE encoders or latent diffusion. By combining a unified vision-language backbone with a pixel-space flow matching head, TUNA-2 supports image understanding, text-to-image generation, and image editing within a single framework. We further instantiated TUNA-2 with both a representation encoder-based variant and an encoder-free monolithic variant, and showed that both designs achieve strong performance across multimodal understanding and generation benchmarks. Our experiments demonstrate that TUNA-2 surpasses prior latent-space unified models such as TUNA and Show-o2 on fine-grained visual understanding benchmarks while remaining competitive with state-of-the-art unified models on image generation. Overall, these results highlight the effectiveness and scalability of pixel-space unified multimodal modelling and suggest a promising direction for future native UMMs.

References

[1] Inclusion AI, Bowen Ma, Cheng Zou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Chenyu Lian, Dandan Zheng, Fudong Wang, Furong Xu, et al. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*, 2025.

[2] 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.

[3] 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.

[4] Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. *CVPR*, 2026.

[5] Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward. *arXiv preprint arXiv:2603.26599*, 2026.

[6] Zhaochong An, Zirui Li, Mingqiao Ye, Feng Qiao, Jiaang Li, Zongwei Wu, Vishal Thengane, Chengzu Li, Lei Li, Luc Van Gool, et al. Video understanding: From geometry and semantics to unified models. *Machine Intelligence Research*, 2026.

[7] Anthropic. Introducing claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7, April 2026. Accessed: 2026-04-24.

[8] 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.

[9] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[10] 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.

[11] 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.

[12] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Fuyu-8b: A multimodal architecture for ai agents, October 2023. URL https://www.adept.ai/blog/fuyu-8b/. Adept blog post.

[13] Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

[14] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

[15] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.

[16] 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.

[17] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[18] Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025.

[19] 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.

[20] Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. Dip: Taming diffusion models in pixel space. *arXiv preprint arXiv:2511.18822*, 2025.

[21] 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.

[22] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

[23] 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.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[25] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *Advances in Neural Information Processing Systems*, 37:52545–52567, 2024.

[26] Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaohua Wang, Hanming Deng, Lewei Lu, Dahua Lin, and Ziwei Liu. From pixels to words–towards native vision-language primitives at scale. *arXiv preprint arXiv:2510.14979*, 2025.

[27] 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.

[28] 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.

[29] 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.

[30] Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, and Ziwei Liu. The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding. *arXiv preprint arXiv:2512.19693*, 2025.

[31] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. Uni-x: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. *arXiv preprint arXiv:2509.24365*, 2025.

[37] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[38] Xin He, Longhui Wei, Jianbo Ouyang, Minghui Liao, Lingxi Xie, and Qi Tian. Emma: Efficient multimodal understanding, generation, and editing with a unified architecture. *arXiv preprint arXiv:2512.04810*, 2025.

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

[40] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

[41] 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.

[42] Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, Taozhi Huang, Jiajia Liu, Qingpei Guo, Ming Yang, Jingdong Chen, and Jun Zhou. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

[43] Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

[44] 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.

[45] 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.

[46] Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20758–20769, 2025.

[47] 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.

[48] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[49] 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.

[50] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[51] Tianle Li, Yongming Rao, Winston Hu, and Yu Cheng. Breen: bridge data-efficient encoder-free multimodal learning with learnable queries. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5384–5395, 2026.

[52] 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*, 2024.

[53] 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.

[54] 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.

[55] 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. URL https://arxiv.org/abs/2505.05422.

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

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

[58] 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.

[59] 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.

[60] Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. *arXiv preprint arXiv:2512.02014*, 2025.

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

[62] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24960–24971, 2025.

[63] 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.

[64] 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.

[65] 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.

[66] Yuwei Niu, Weiyang Jin, Jiaqi Liao, Chaoran Feng, Peng Jin, Bin Lin, Zongjian Li, Bin Zhu, Weihao Yu, and Li Yuan. Does understanding inform generation in unified multimodal models? from analysis to path forward. *arXiv preprint arXiv:2511.20561*, 2025.

[67] OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, March 2026. Accessed: 2026-04-24.

[68] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3170–3180, 2023.

[69] Haonan Qiu, Shikun Liu, Zijian Zhou, Zhaochong An, Weiming Ren, Zhiheng Liu, Jonas Schult, Sen He, Shoufa Chen, Yuren Cong, et al. Histream: Efficient high-resolution video generation via redundancy-eliminated streaming. *arXiv preprint arXiv:2512.21338*, 2025.

[70] 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.

[71] A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint*, 2024.

[72] 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.

[73] Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21197–21208, 2025.

[74] 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.

[75] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[76] Tao Shen, Xin Wan, Taicai Chen, Rui Zhang, Junwen Pan, Dawei Lu, Fanding Lei, Zhilin Lu, Yunfei Yang, Chen Cheng, et al. Mammothmoda2: A unified ar-diffusion framework for multimodal understanding and generation. *arXiv preprint arXiv:2511.18262*, 2025.

[77] Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, et al. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.

[78] 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.

[79] 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.

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

[81] Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

[82] Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

[83] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

[84] Shengbang Tong, David Fan, John Nguyen, Ellis Brown, Gaoyue Zhou, Shengyi Qian, Boyang Zheng, Théophane Vallaeys, Junlin Han, Rob Fergus, et al. Beyond language modeling: An exploration of multimodal pretraining. *arXiv preprint arXiv:2603.03276*, 2026.

[85] 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.

[86] 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.

[87] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

[88] Han Wang, Yongjie Ye, Bingru Li, Yuxiang Nie, Jinghui Lu, Jingqun Tang, Yanjie Wang, and Can Huang. Vision as lora. *arXiv preprint arXiv:2503.20680*, 2025.

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

[90] 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.

[91] Xiang Wang, Zhifei Zhang, He Zhang, Zhe Lin, Yuqian Zhou, Qing Liu, Shiwei Zhang, Yijun Li, Shaoteng Liu, Haitian Zheng, et al. Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. *arXiv preprint arXiv:2511.20520*, 2025.

[92] 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.

[93] Zeyu Wang, Zilong Chen, Chenhui Gou, Feng Li, Chaorui Deng, Deyao Zhu, Kunchang Li, Weihao Yu, Haoqin Tu, Haoqi Fan, et al. Lightfusion: A light-weighted, double fusion framework for unified multimodal understanding and generation. *arXiv preprint arXiv:2510.22946*, 2025.

[94] 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.

[95] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

[96] Hongyang Wei, Baixin Xu, Hongbo Liu, Size Wu, Jie Liu, Yi Peng, Peiyu Wang, Zexiang Liu, Jingwen He, Yidan Xietian, et al. Skywork unipic 2.0: Building kontext model with online rl for unified multimodal model. *arXiv preprint arXiv:2509.04548*, 2025.

[97] 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.

[98] 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.

[99] 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.

[100] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[101] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025.

[102] 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.

[103] 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.

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

[105] 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.

[106] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025.

[107] 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.

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

[109] 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.

[110] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

[111] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.

[112] 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.

[113] Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good tokenizers. In *The Fourteenth International Conference on Learning Representations*, 2025.

[114] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025.

[115] Yuhang Yang, Fan Zhang, Huaijin Pi, Shuai Guo, Guowei Xu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Gloria: Consistent character video generation via content anchors. *arXiv preprint arXiv:2603.29931*, 2026.

[116] 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.

[117] Zebin You, Xiaolu Zhang, Jun Zhou, Chongxuan Li, and Ji-Rong Wen. Llada-o: An effective and length-adaptive omni diffusion model. *arXiv preprint arXiv:2603.01068*, 2026.

[118] 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.

[119] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[120] Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. *arXiv preprint arXiv:2511.20645*, 2025.

[121] 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.

[122] 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.

[123] Boqiang Zhang, Lei Ke, Ruihan Yang, Qi Gao, Tianyuan Qu, Rossell Chen, Dong Yu, et al. Penguin-vl: Exploring the efficiency limits of vlm with llm-based vision encoders. *arXiv preprint arXiv:2603.06569*, 2026.

[124] Huichao Zhang, Liao Qu, Yiheng Liu, Hang Chen, Yangyang Song, Yongsheng Dong, Shikun Sun, Xian Li, Xu Wang, Yi Jiang, et al. Nextflow: Unified sequential modeling activates multimodal understanding and generation. *arXiv preprint arXiv:2601.02204*, 2026.

[125] Letian Zhang, Sucheng Ren, Yanqing Liu, Xianhang Li, Zeyu Wang, Yuyin Zhou, Huaxiu Yao, Zeyu Zheng, Weili Nie, Guilin Liu, et al. Openvision 3: A family of unified visual encoder for both understanding and generation. *arXiv preprint arXiv:2601.15369*, 2026.

[126] 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.

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

[128] 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.

[129] Zijian Zhou, Shikun Liu, Haozhe Liu, Haonan Qiu, Zhaochong An, Weiming Ren, Zhiheng Liu, Xiaoke Huang, Kam Woh Ng, Tian Xie, et al. Scaling zero-shot reference-to-video generation. *arXiv preprint arXiv:2512.06905*, 2025.

Appendix

A Additional Related Work

A.1 Encoder-Free Multimodal Understanding Models

Large multimodal models (LMMs) have traditionally adopted a modular design, combining pixel-space representation encoders with LLM decoders. Early work on LMMs explored integration strategies between representation encoders and LLM decoders, such as using cross-attention in Flamingo [2] and MLP connectors in LLaVA [57]. Subsequent research largely followed LLaVA and scaled this paradigm to larger models [89, 10, 90, 6], datasets [126, 97, 3], and diverse understanding tasks [21, 52, 73]. More recently, an alternative design of employing a monolithic, encoder-free transformer to natively process vision and language has emerged [88, 51, 46]. Fuyu [12], EVE [25], Chameleon [80], Mono-InternVL [62] and NEO [26] employ simple MLP or patch embedding layers to tokenize raw image pixels into patches, and jointly process these image patches with language tokens using a single transformer. In this work, we show that both the representation encoder-based design and the monolithic design can be integrated into pixel-space UMMs and achieve high performance on multimodal understanding.

A.2 Encoder-Free Visual Generation Models

Recent visual generation models [115, 17, 4, 5, 129, 69] typically operate in a compressed latent space using KL- or VQ-regularized VAEs [74, 28, 98, 27, 78]. Although pixel-space diffusion or flow matching is often considered more challenging, recent work such as PixelFlow [18], DiP [20], PixelDiT [120] and JiT [50] increasingly suggests that pixel-space flow matching has the potential to match or surpass latent diffusion models. However, these studies are usually limited to small-scale settings (e.g., class-conditioned generation on ImageNet [75]). We demonstrate in TUNA-2 that pixel-space flow matching can be scaled up to large-scale unified multimodal pretraining and support free-form text-to-image generation and image editing.

B Additional Experimental Results: Image Reconstruction

Table 6: Image reconstruction performance for different visual tokenizers.

Tokenizer	Res.	rFID↓	PSNR↑	SSIM↑
Specialized tokenizers				
SD-VAE	256	1.06	28.62	0.86
GigaTok	256	0.51	21.32	0.69
VA-VAE	256	0.26	28.59	0.80
DC-AE	512	0.22	26.15	0.71
MAE-Tok	512	0.62	–	–
TexTok	512	0.73	24.45	0.66
FLUX.1[dev]-VAE†	512	0.06	33.65	0.93
Unified tokenizers				
UniTok	256	0.38	–	–
TokenFlow	384	0.63	22.77	0.73
X-Omni†	512	8.30	15.66	0.38
MingTok†	512	0.53	23.49	0.61
RAE	256	0.61	19.20	0.44
PS-VAE	256	0.20	28.79	0.82
TUNA-R	512	**0.12**	32.22	**0.93**
TUNA-2	512	0.15	**32.80**	**0.93**

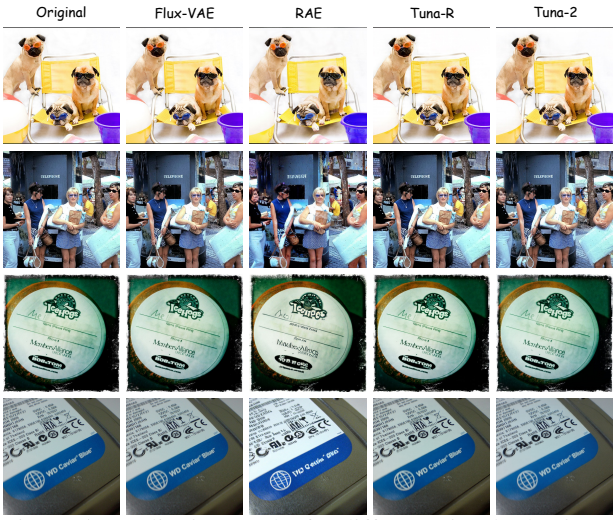

Figure 5: Qualitative results for different visual tokenizers.

To further investigate the image generation capability of TUNA-2, we examine whether the model can faithfully reconstruct an input image from its corresponding pixel-space visual representations. To this end, we perform lightweight finetuning on an image reconstruction task and evaluate reconstruction quality on the ImageNet validation set [24]. As shown in Table 6, both TUNA-R and TUNA-2 achieve strong reconstruction performance, ranking first among unified tokenizers and approaching specialized

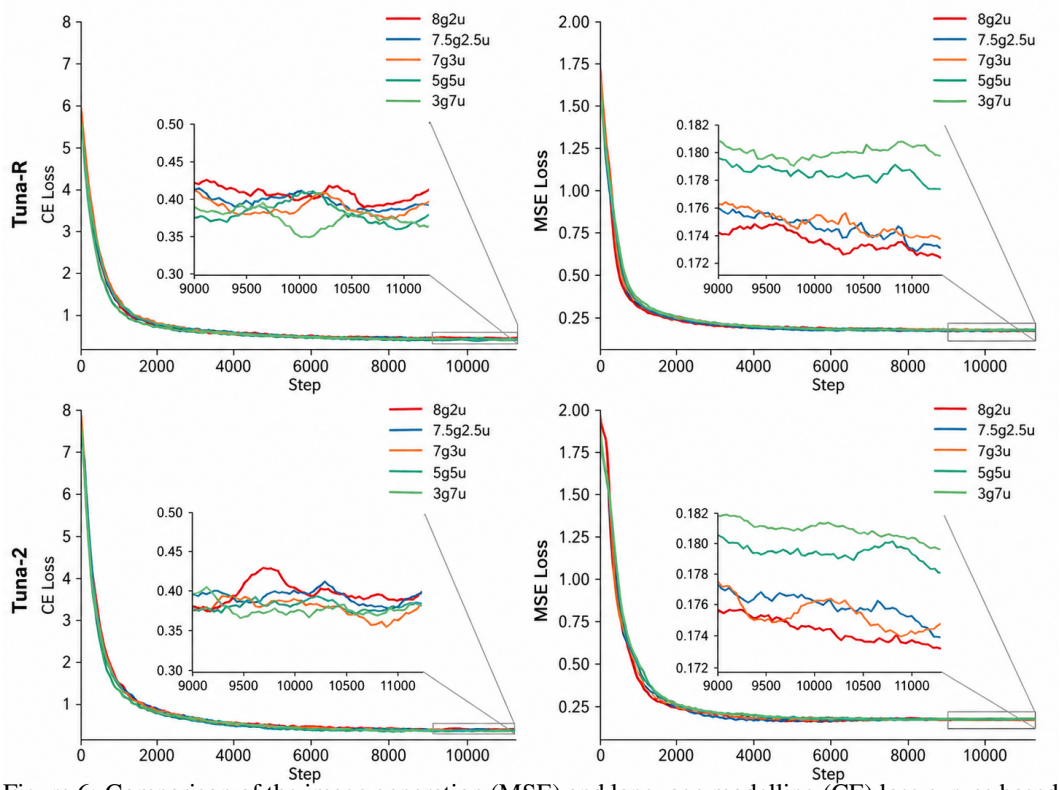

Figure 6: Comparison of the image generation (MSE) and language modelling (CE) loss curves based on different understanding-generation data ratios during model training.

image tokenizers such as the VAE model in FLUX.1 [dev] [11]. Figure 5 further shows that our models produce substantially better reconstructions than other non-KL-regularized VAE approaches such as RAE [127]. These results indicate that our pixel-space unified visual representation can support strong image reconstruction and generation quality even without relying on VAE models.

C Ablation: Model Training Dynamics

To better understand the learning dynamics of our joint multimodal training paradigm, we conduct a series of pretraining ablation studies on both TUNA-R and TUNA-2 by varying the sampling ratio between generation and understanding data. We use the notation `xgyu` (e.g., `8g2u`) to represent a generation-to-understanding sampling ratio of $x : y$. As shown in Figure 6, increasing the proportion of either generation or understanding data consistently reduces its corresponding training loss, namely MSE for flow matching and cross entropy loss (CE) for language modelling. Notably, the MSE loss is more sensitive to changes in the sampling ratio, while the CE loss varies within a relatively smaller range across different ratios. This suggests that both tasks benefit from scaling up the amount of their corresponding training data, with the generation objective being more affected by the data mixture. We also observe that a generation-to-understanding ratio of $7 : 3$ (`7g3u`) achieves the best trade-off between the two objectives, yielding a strong balance between generation and understanding performance. We therefore adopt this data sampling ratio in all experiments.

D Analysis: Attention Map Visualizations

To qualitatively analyze and compare the cross-modal alignment capabilities learned by different models, we visualize the attention maps of selected keywords in the input prompt with respect to the input image. We focus on the TUNA model family, including TUNA, TUNA-R, and TUNA-2, and compare them with representative encoder-based open-source LMMs, including models with

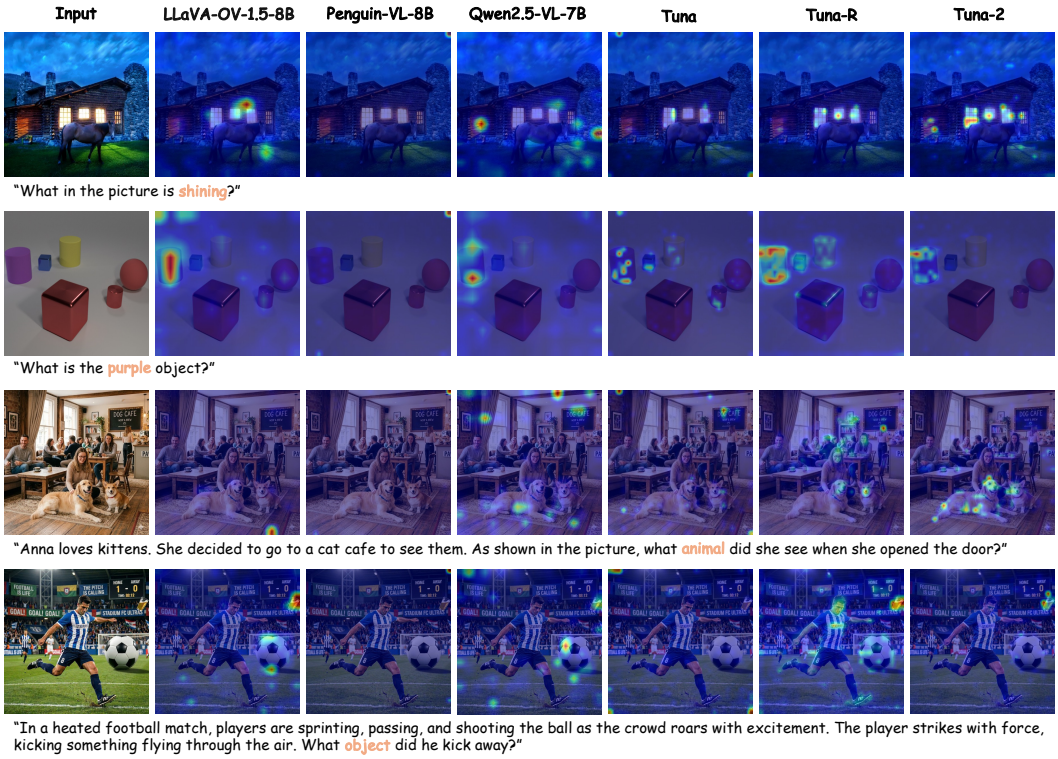

Figure 7: Attention map visualization for TUNA-R, TUNA-2 and other baseline models. Red area denotes high attention scores and Blue area denotes low attention scores.

ViT-based encoders, such as LLaVA-OneVision-1.5 [3] and Qwen2.5-VL [10], as well as models with LLM-based encoders, such as Penguin-VL [123]. The results are shown in Figure 7.

Our first observation is that TUNA-2 exhibits more accurate vision-language alignment in basic perception scenarios. For example, in the "shining window" case, TUNA-2 consistently highlights the regions that are semantically associated with "shining", while other models tend to provide only coarse or incomplete localization. In the "purple object" case, the attention activations of TUNA-2 are well aligned with the target object, whereas other methods often show dispersed attention or spurious activations in irrelevant regions.

We further construct examples with misleading linguistic contexts to evaluate whether models overly rely on language priors. For example, in the "dog cafe" case, the prompt suggests a cat cafe scenario, whereas the actual animals in the image are dogs. Some models tend to rely on textual cues within the image (e.g. "dog cafe" sign) instead of the actual visual subjects. In contrast, TUNA-2 places its attention on the truly relevant visual regions, demonstrating more robust cross-modal alignment under misleading linguistic contexts.

In the more challenging counterintuitive "football match" case, we introduce both strong linguistic priors and salient visual distractors. Specifically, the prompt contains strong football-related cues, such as "football match", "players", and "kicking". The image also includes a large and visually salient football as a distractor. However, the actual object being kicked by the player is a glass cup. Most models are easily misled by either the textual prior or the salient distractor and therefore attend to incorrect regions. In contrast, TUNA-2 accurately localizes the key object that is consistent with the question semantics, showing stronger robustness in such counterintuitive settings.

Overall, as an encoder-free native multimodal architecture, TUNA-2 produces more accurate and stable attention distributions across all cases. These results suggest that TUNA-2 learns more reliable visual representations, leading to more consistent cross-modal alignment and better robustness to misleading language priors and visual distractors.

E Compute Resource

All experiments were conducted on an internal GPU cluster equipped with NVIDIA H100 GPUs (80GB memory). For the main pretraining experiments, we used 64 nodes with 8 GPUs per node (512 GPUs in total). Stage 1 pretraining was performed on approximately 3T multimodal and text tokens for 300k steps, requiring roughly 25 days of training. Stage 2 supervised finetuning (SFT) was conducted on approximately 0.5T tokens for 50k steps and required approximately 5 additional days of training. All training stages used a maximum sequence length of 16K tokens per GPU. For TUNA-R, the additional connector alignment stage was performed on the same hardware configuration for 3k steps. For the 1.5B-scale ablation studies in Section 3.3, we used 16 nodes for all experiments. The training ratio ablation experiments were conducted for 20k steps, while the masking-based feature learning ablations were trained for 100k steps.

F Limitations

While TUNA-2 demonstrates strong performance as an encoder-free unified multimodal model, several limitations remain. First, our current study is restricted to image understanding and generation. Existing video generation models still rely heavily on temporal compression in pretrained VAEs, and removing the VAE also removes this advantage, leaving efficient pixel-space video generation as an open research problem. Second, to ensure a fair comparison with prior work, we adopt Qwen2.5-7B-Instruct as the base LLM decoder, which may not represent the strongest available backbone. We expect that replacing it with a newer and more capable base model (e.g. Qwen3 [112]) could further improve performance.

G Broader Impact

TUNA-2 advances unified multimodal modelling by showing that visual understanding and generation can be learned directly from pixel embeddings without relying on pretrained vision encoders. This simplified encoder-free design may make future multimodal systems easier to scale up, adapt, and extend, while improving fine-grained visual perception for applications such as visual assistance, document understanding, creative content generation, and image editing. At the same time, stronger unified models that combine understanding and generation also raise potential risks, including the creation or manipulation of misleading visual content, unintended biases inherited from large-scale training data, and possible misuse in settings that require faithful visual evidence. We hope our findings encourage further research toward more capable, efficient, and responsible native unified multimodal systems.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have justified our claims in the abstract and introduction through extensive experiments and analysis in Section 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included some discussions of our method's limitations in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: There are no theorems and proofs in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included necessary details of our experiment settings in Section 3.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We did not include access to data and code in the submission; However, we will release code and model, pending approval.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided detailed hyperparameter settings in Section 3.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We did not report statistical significance because it would be too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details could be seen in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research fully complies with NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have included some discussions on the broader impacts of this research in Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: Our released code will contain information on usage guidelines and restrictions.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The open-source datasets that we used for model training are discussed and cited in Section 3.1.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.