Lance: Unified Multimodal Modeling by Multi-Task Synergy

Fengyi Fu[1*‡] **Mengqi Huang**[1*†‡] **Shaojin Wu**[1*] **Yunsheng Jiang**[1*] **Yufei Huo**[1‡] **Hao Li**[1] **Yinghang Song**[1] **Fei Ding**[1] **Jianzhu Guo**[1†§] **Qian He**[1] **Zheren Fu** **Zhendong Mao** **Yongdong Zhang**

[1]Intelligent Creation Lab, ByteDance

[*]Equal contribution, [†]Corresponding Author, [§]Project lead, [‡]Work was done during their internship.

Abstract

We present **Lance**, a lightweight native unified model supporting multimodal understanding, generation, and editing for both images and videos. Rather than relying on model capacity scaling or text-image-dominant designs, Lance explores a practical paradigm for unified multimodal modeling via collaborative multi-task training. It is grounded in two core principles: unified context modeling and decoupled capability pathways. Specifically, Lance is trained from scratch and employs a dual-stream mixture-of-experts architecture on shared interleaved multimodal sequences, enabling joint context learning while decoupling the pathways for understanding and generation. We further introduce modality-aware rotary positional encoding to mitigate interference among heterogeneous visual tokens and boost cross-task alignment. During training, Lance adopts a staged multi-task training paradigm with capability-oriented objectives and adaptive data scheduling to strengthen both semantic comprehension and visual generation performance. Experimental results demonstrate that Lance substantially outperforms existing open-source unified models in image and video generation, while retaining strong multimodal understanding capabilities.

Date: May 15, 2026
Correspondence: huangmengqi.98@bytedance.com, xiaoyu.e@bytedance.com
Project Page: https://lance-project.github.io

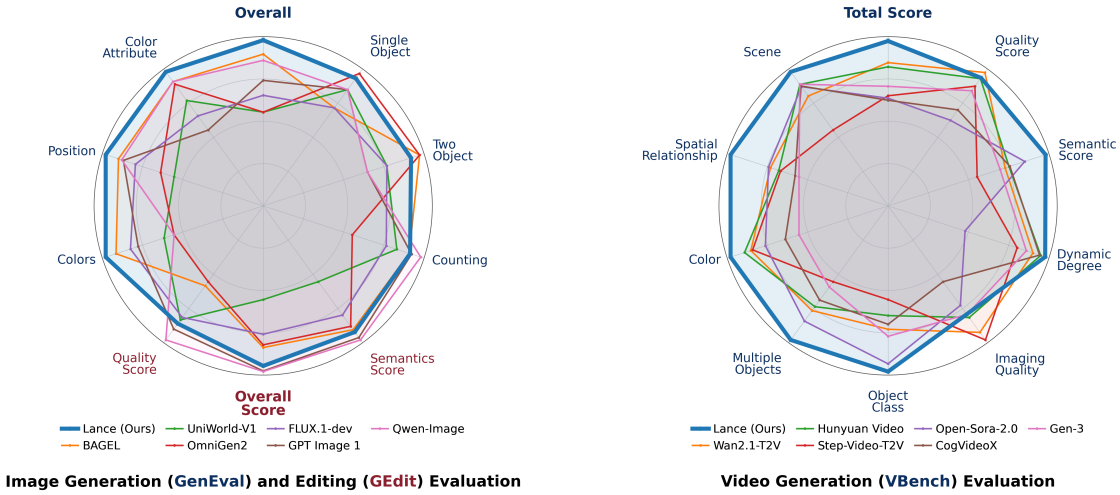


Figure 1 Comparison of Lance against representative baselines on multimodal benchmarks.

Paradigm	Method	UND. (Image to Text)			UND. (Video to Text)			GEN. (Image)			GEN. (Video)				Emergent Generalization
		Cap.	Per.	Rea.	Cap.	Per.	Rea.	T2I	Edit	S2I	T2V	I2V	Edit	S2V	
Non-native Unified	MetaQuery-XL [88]	✓	✓	✓				✓		✓					
	SEED-X [32]	✓	✓	✓				✓	✓						
	TokenFlow-XL [92]	✓	✓	✓				✓							
	ILLUME [110]	✓	✓	✓				✓	✓						
	InternVL-U [103]	✓	✓	✓				✓	✓						
	UniVideo [119]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Native Unified	Chameleon [100]	✓	✓	✓				✓							
	LWM [69]	✓	✓	✓	✓	✓	✓	✓			✓				
	Janus [123]	✓	✓	✓				✓							
	Janus-Pro [14]	✓	✓	✓				✓							
	Transfusion [149]	✓	✓	✓				✓							
	Emu3 [115]	✓	✓	✓	△	△	△	✓			✓				
	Show-o [133]	✓	✓	✓				✓	✓						
	Show-o2 [134]	✓	✓	✓	✓	✓	✓	✓	✓		△				
	Bagel [22]	✓	✓	✓				✓	✓	✓					✓
	Mogao [63]	✓	✓	✓				✓	△	△					
	HaploOmni [132]	✓	✓	✓	✓	✓	✓	✓			✓				
	VILA-U [130]	✓	✓	✓	✓	✓	✓	✓			✓				
	HunyuanImage 3.0 [8]	△	△	△				✓	✓						
	Emu3.5 [19]	✓	✓	✓	△	△	△	✓	✓	△	△	△			✓
	TUNA [77]	✓	✓	✓	✓	✓	✓	✓	✓		✓				
	TUNA-2 [78]	✓	✓	✓				✓	✓						
	Lance (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1 Comparison of multimodal unified models by supported task categories. ✓ indicates explicit support; △ indicates description-only support without official code; blank cells indicate no explicit report. Cap., Per., Rea. indicate understanding ability on captioning, perception, and reasoning. The last column denotes whether the model exhibits emergent generalization on unseen tasks. Models are categorized as native or non-native unified models based on whether they are jointly pre-trained as a unified architecture or assembled from separately pre-trained components.

1 Introduction

Multimodal artificial intelligence is increasingly moving toward a native unified paradigm, where understanding, reasoning, and generation are integrated within a unified framework. Recently, large language models [2, 4, 5, 16, 56, 70] have driven rapid advances in image and video understanding, while diffusion- and flow-based models [25, 40, 53, 54, 67, 97, 98, 140] have advanced high-fidelity image and video generation. However, most existing systems still evolve along two separate paths: understanding models emphasize semantic reasoning and instruction following, while generative models focus on visual synthesis and spatiotemporal dynamics. Unifying these capabilities in a single unified model remains a central challenge in developing multimodal foundation models with greater generality and stronger practical utility.

Recent unified multimodal models [19, 22, 63, 77, 100, 134] have made encouraging progress, yet two fundamental limitations remain. First, the visual-representation requirements of understanding and generation are inherently misaligned: the former benefits from high-level semantic features aligned with language, whereas the latter requires low-level continuous representations that preserve texture, geometry, and temporal dynamics. Existing approaches therefore typically follow one of two directions. One line of work [19, 77, 100, 115, 133] attempts to support both tasks with a unified visual representation, yielding a simpler modeling formulation but often struggling to balance semantic reasoning and generation quality. Another line [22, 63, 134] adopts decoupled semantic and generative representations, alleviating representational mismatch at the cost of increased architectural and optimization complexity.

Second, and more importantly, existing unified models remain limited in task coverage and training formulation. As summarized in Table 1, most prior methods [32, 69, 92, 100, 123] are still largely confined to text-image domains or partial task combinations, leaving the full image-video understanding and generation space insufficiently explored. Although recent unified models [22, 77, 134] have progressively extended to the video domain, they typically cover only limited subsets of the full image-video task space, while diverse generation-oriented tasks such as editing and subject-driven generation are often introduced as downstream fine-tuning skills rather than being systematically optimized within a unified multi-task training process. Meanwhile, the comparison in Table 1 further suggests that models with broader task coverage are more likely to exhibit emergent generalization on unseen tasks. This motivates us to view multi-task learning not simply

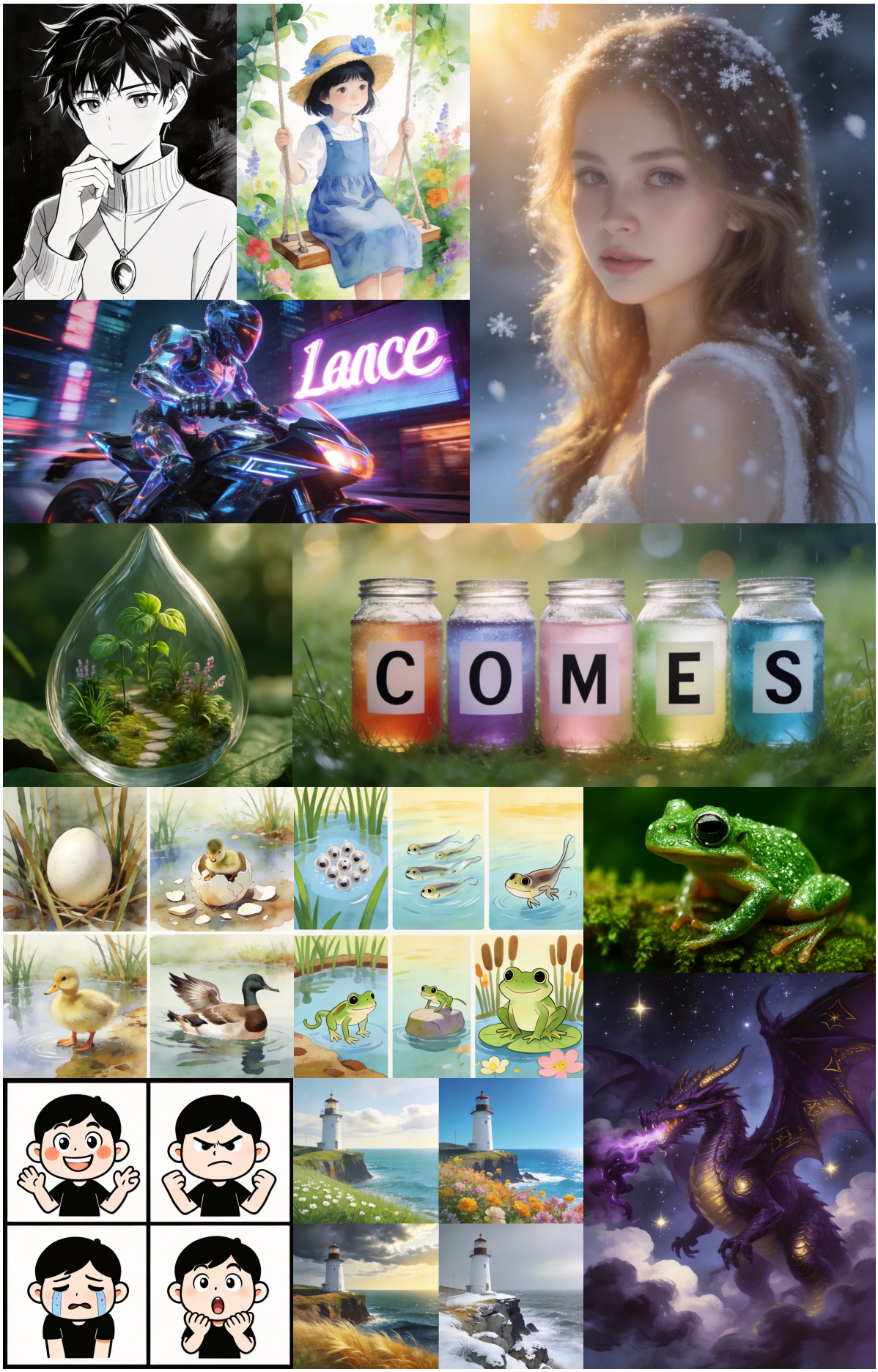


Figure 2 Text-to-image generation (T2I) with Lance.

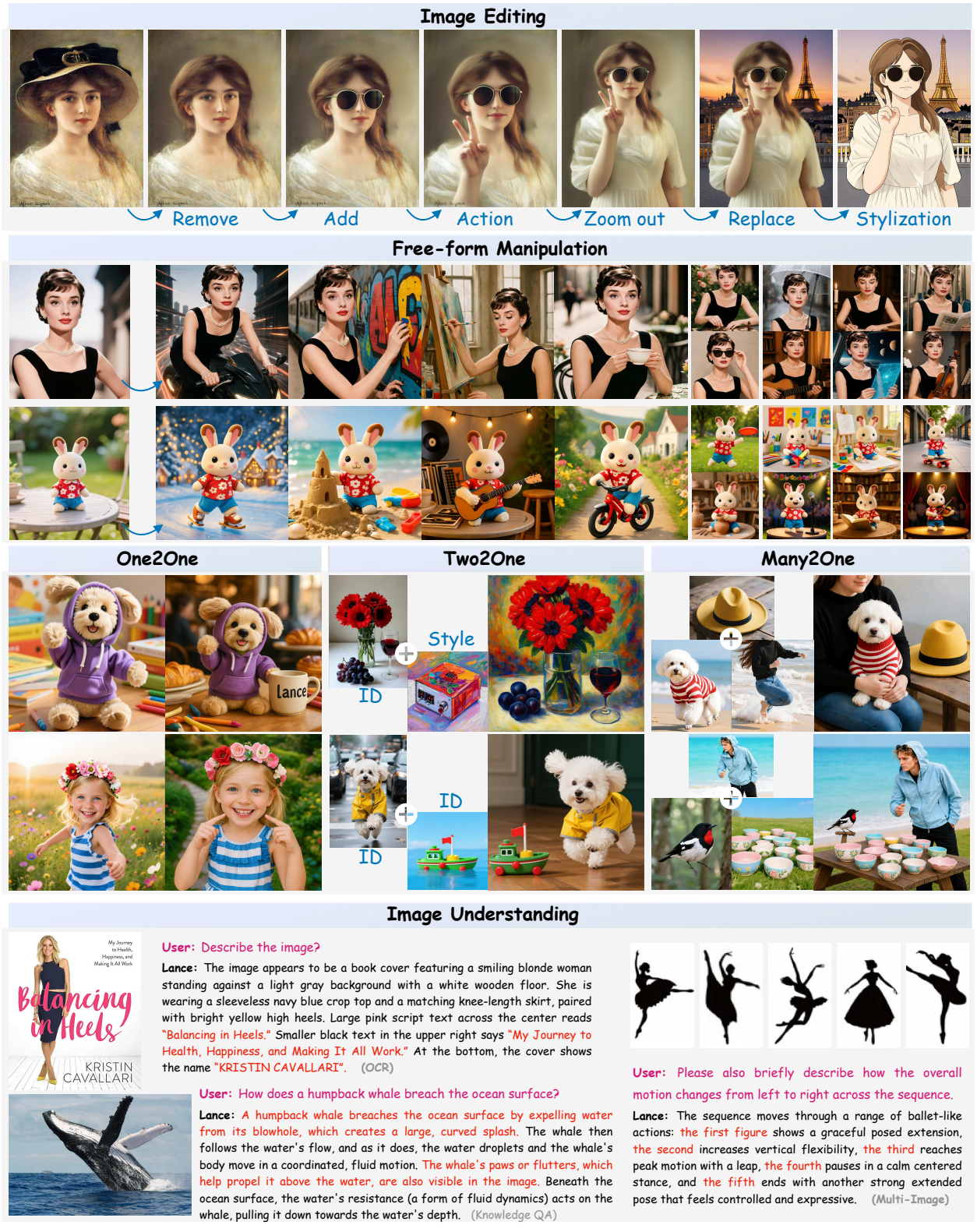


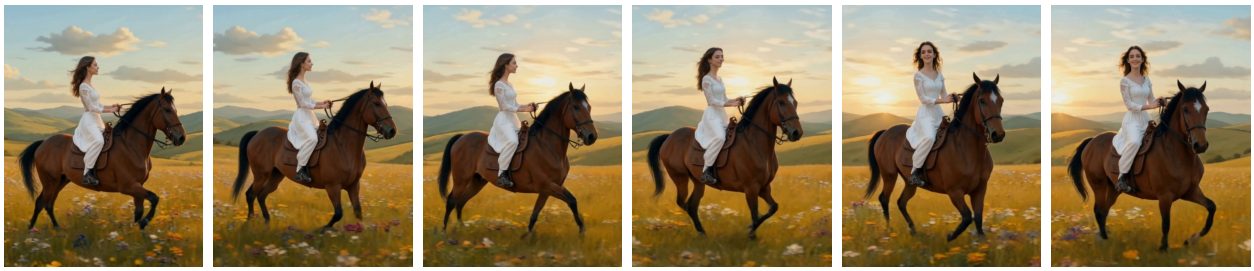
Figure 3 Any-to-image generation (X2I) and image understanding (I2T) with Lance.

General Text-to-Video Generation

A stylish tabby cat wearing a black top hat paints on a canvas in a cozy sunlit art studio.

A panda and a humanoid robot are boxing in a grand palace courtyard.

A handsome young man playing a guitar in an elegant sunlit conservatory hall. He first looks down at the guitar with focused concentration, then lifts his head toward the camera and reveals a gentle smile.



Vertical portrait shot, a fashion-forward woman riding a horse across a meadow with wildflowers, rolling hills, and luminous bright morning light. Camera side shot. The woman facing forward first, then turning her head to face the camera with a smile.

Physics-Aware Text-to-Video Generation

A red balloon and a black metal ball fall from the same height on Earth.

A red balloon and a black metal ball fall from the same height on Moon.

Pour clear water into a transparent glass cup.

Pour fine sand into a transparent glass cup.

A sheet of tissue paper burns on a shallow iron plate.

A smooth ball and a sticky ball slide down a slope.

Figure 4 Text-to-video generation (T2V) with Lance.

Figure 5 Any-to-video generation (X2V) and video understanding (V2T) with Lance.

as capability aggregation, but as a way to promote transfer across modalities and task formulations.

Based on this observation, we present **Lance**, a lightweight native unified multimodal model that systematically integrates joint learning across X2T, X2I, and X2V tasks, covering image and video understanding, generation, and editing within a single framework. By unifying these task families in a single native model, Lance aims to better harness cross-task synergy and further advance the potential of unified multimodal modeling. Lance is designed to balance *unified context modeling* with *decoupled capability pathways* from both the architectural and training perspectives. Architecturally, it adopts a shared interleaved multimodal sequence representation to enable unified context learning, while employing a dual-stream mixture-of-experts framework to allocate dedicated capacity to semantic reasoning and visual synthesis. To better coordinate heterogeneous visual tokens within the unified context sequence, we further introduce modality-aware rotary positional encoding, MaPE, which mitigates positional interference and improves cross-task contextual alignment. In terms of training, Lance follows a staged multi-task training paradigm that casts diverse understanding, generation, and editing tasks into a unified task formulation, and combines capability-oriented objectives with adaptive data scheduling to progressively strengthen semantic understanding and visual synthesis.

Extensive experiments show that Lance achieves strong performance across multimodal understanding and generation benchmarks, with qualitative examples shown in Figures 2 to 5. With only 3B activated parameters, Lance substantially outperforms existing open-source unified models on image and video generation tasks as shown in Figure 1, while maintaining advanced multimodal understanding ability. Notably, all these gains are achieved within a 128-GPU training budget, highlighting the feasibility of resource-efficient unified multimodal modeling.

Our main contributions are summarized as follows:

(1) **Concepts:** We present Lance, a lightweight native unified multimodal model that explicitly supports the full spectrum of image/video understanding and generation tasks within a single model, extending unified modeling beyond text-image domains and partial task coverage. Lance emphasizes multi-task synergy not as simple capability aggregation, but as a mechanism for promoting transfer across modality-task boundaries.

(2) **Technique:** We develop a dual-stream mixture-of-experts architecture that preserves a shared interleaved multimodal sequence representation while allocating dedicated visual representations and model capacity to understanding and generation. We further introduce a modality-aware positional encoding scheme and a staged multi-task training paradigm to improve heterogeneous visual token coordination and cross-task context modeling.

(3) **Performance:** Extensive experiments demonstrate that Lance achieves competitive performance across multimodal understanding and generation benchmarks with only 3B activated parameters.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) have become the dominant paradigm for image and video understanding by aligning pretrained visual encoders with powerful language backbones. Representative early systems include Flamingo [2], IDEFICS [55], and InstructBLIP [20], while later open-source families such as LLaVA [56, 70–72], Qwen-VL [3–5, 113], and InternVL [15, 16, 31, 114] further improve instruction following, high-resolution perception, and long-context multimodal reasoning. This line of work mainly follows the LLaVA paradigm [70], in which visual inputs are first encoded by a vision encoder [93, 107] and then concatenated with text tokens for joint modeling by a language model decoder. Some proprietary models such as GPT [1] and Gemini [101, 102] also demonstrate strong multimodal reasoning ability. Recent progress further extends these models to interleaved image-text modeling [19, 22, 138] and video understanding [60, 64, 139]. Despite their strong semantic abstraction and cross-modal alignment capabilities, these models are primarily optimized for understanding and text generation, rather than native visual synthesis.

2.2 Visual Generative Models

Visual generation has been dominated by diffusion- and flow-based frameworks [25, 29, 30, 39, 44, 53, 54, 67, 82, 85, 125], which serve as mainstream paradigms for high-fidelity image and video synthesis. As for image

generation, representative large-scale systems include Stable Diffusion [25, 91, 95, 129], FLUX [53, 54], Qwen-Image [122], and HunyuanImage 3.0 [8], while multimodal image generation models such as RealCustom++ [44, 83] and UNO series [17, 126, 127] further advance these frameworks by supporting diverse multimodal conditional inputs. As for video generation, recent systems such as Wan [109], HunyuanVideo [121] and CogVideo [40, 140] demonstrate the effectiveness of continuous latent modeling with dedicated temporal VAEs. In contrast to continuous latent generators, autoregressive visual token models [9, 24, 43, 51, 84, 89, 94, 104] formulate image generation as next-token prediction, providing a simpler unified token interface, but often face trade-offs in visual fidelity and generation efficiency. Recently, several studies [26, 61, 73] have explored hybrid frameworks that combine diffusion modeling with autoregressive modeling, aiming to leverage the advantages of both in generation quality and modeling flexibility, thereby further advancing visual generation capabilities.

2.3 Unified Multimodal Models

Recent unified multimodal models (UMMs) attempt to bridge multimodal understanding and visual generation within a single framework. One line follows a fully autoregressive formulation, represented by Chameleon [100], Emu3/Emu3.5 [19, 115], and more recent systems such as TokenFlow [92], HunyuanImage 3.0 [8]. These models cast both understanding and generation into next-token prediction under a shared token space. These models offer a clean unified interface and naturally support mixed-modality sequence modeling, but they may still face nontrivial trade-offs among reasoning ability, visual fidelity, and generation efficiency. Another line adopts autoregressive–diffusion hybrid formulations, combining language modeling for text with diffusion- or flow-based modeling for visual generation. Representative works include Transfusion [149], Show-o/Show-o2 [133, 134], BLIP3-o [11], BAGEL [22], and others [21, 28, 37, 58, 77, 80, 105, 111, 146]. Within this family, recent work further explores decoupling in representation design, module architecture, and optimization. For instance, Janus-series models [80, 146] decouple visual encoding for understanding and generation; RealGeneral [66] tames a pretrained video foundation model for unified image generation and editing; Show-o2 [134] integrates autoregressive language modeling with flow matching, extending native unification to both image and video modalities; BAGEL [22] studies expert specialization under a shared decoder-only backbone; TUNA [77] emphasizes unified continuous visual representations; and InternVL-U [103] couples a strong open MLLM with a specialized generation head. In addition to native unified models, modular bridging systems such as OmniBridge [131] connect pretrained understanding and generation models through latent-space alignment, offering a more lightweight but less fully native alternative.

Although unified multimodal modeling has advanced rapidly, much of the literature remains image-centric. Extending unified modeling to the video domain is substantially more challenging because it requires not only semantic understanding but also temporal reasoning, motion modeling, long-context generation, and consistent editing. Early general any-to-any or modular systems such as NEXT-GPT [128] and GPT4Video [118] extend MLLMs with external generative backends to support multimodal understanding and video generation, but their video synthesis capability is still largely mediated through additional generators rather than native joint video modeling. More recent video-focused frameworks, including Omni-Video [99], UniVideo [119], and TV2TV [36], move closer to genuinely unified video models by jointly addressing video understanding, generation, editing, or interleaved language-video modeling under a more integrated architecture. Meanwhile, several task-unified video editing frameworks, such as AnyV2V [52], VACE [47], UNIC [141], EditVerse [48], and FullDiT [49], expand the controllability of video generation, but typically do not aim for full understanding-generation unification within a single multimodal model. Overall, multi-task synergy for image-video unified multimodal modeling remains to be further explored.

3 Methodology

The core idea of Lance is that broad multi-task learning can further unlock the potential of unified multimodal models. However, different task families, such as multimodal understanding, generation, and editing, impose substantially different requirements on modeling objectives, visual representations, and optimization dynamics. An effective unified model should therefore enable different tasks to interact within *unified context learning*, while mitigating interference among heterogeneous objectives through *decoupled capability pathways*.

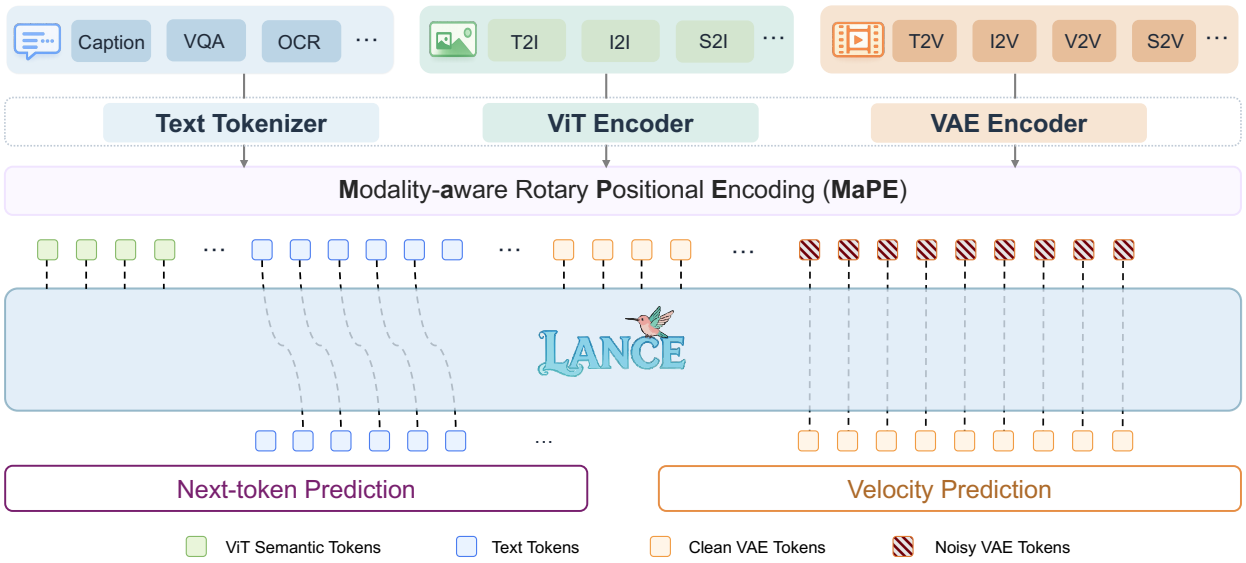


Figure 6 Overview of Lance. Given multi-task inputs spanning X2T, X2I, and X2V, Lance encodes all input tokens into a unified MaPE-enhanced multimodal context sequence. The dual-expert backbone performs generalized 3D causal attention over the shared context and produces task-specific hidden states, which are further decoded by an LM head for autoregressive next-token prediction and by a flow head for velocity prediction in the visual latent space.

3.1 Design Motivation and Principles

Lance is built upon two principles: *unified context learning* and *decoupled capability pathways*. Unified context learning is enabled by interleaved multimodal sequence modeling and multi-task collaborative optimization, while decoupled capability pathways are motivated by the following observations.

Autoregressive vs. Diffusion. Autoregressive next-token prediction remains the dominant paradigm for language modeling [1, 68, 106] and multimodal understanding [4, 60, 136]. In contrast, high-quality image and video synthesis is more effectively modeled in continuous latent spaces with diffusion or flow-matching objectives [7, 23, 54, 57, 122]. Some unified models [92, 100, 115, 130] also explore fully autoregressive formulations for joint understanding and generation, which may suffer from sequential decoding and limited generation efficiency. We therefore adopt autoregressive language modeling for understanding and flow matching for generation.

Unified Visual Representations vs. Decoupled Visual Representations. Understanding and generation rely on different forms of visual information. Understanding mainly benefits from high-level semantic visual features that are well aligned with language (*e.g.*, SigLIP 2 [107] or Qwen2.5-VL [5]), whereas generation relies on low-level latent representations that preserve appearance and spatiotemporal structure [109]. Some existing works [77] have explored shared visual representations, but a single representation may be insufficient to simultaneously satisfy semantic reasoning and high-fidelity synthesis. Meanwhile, recent studies [142, 147] suggest that semantic features can also benefit generation modeling. Lance therefore keeps semantic visual tokens and generative latent tokens decoupled, while organizing them within a shared interleaved multimodal sequence for unified context learning.

Shared Backbone vs. Specialized Expert Capacity. A fully shared backbone that uses single stream to process various modalities [42, 77, 134] offers a clean unified architecture, but it forces understanding and generation to compete for the same parameters under substantially different objectives. Recent evidence from Bagel [22] and HunyuanImage 3.0 [8] further suggests that decoupling generation-oriented parameters and understanding-oriented parameters yields clear advantages over dense shared backbones. These observations motivate Lance to preserve a unified multimodal token interface for bottleneck-free context fusion, while allocating specialized expert capacity to understanding and generation pathways.

3.2 Overall Architecture

Overall Framework. An overview of our framework is shown in Figure 6. Given interleaved inputs of text, images, and videos, Lance first converts each modality into task-appropriate token representations. These heterogeneous tokens are then organized into a shared interleaved multimodal sequence with modality-aware rotary positional encoding, supporting unified context modeling across diverse task formats. To reconcile unified context learning with task-specific capability specialization, Lance adopts a dual-expert architecture initialized from Qwen2.5-VL [5]. The understanding expert, denoted as LLM_{UND}, processes text and semantic visual tokens for multimodal reasoning and text generation, while the generation expert, denoted as LLM_{GEN}, processes VAE latent tokens for visual synthesis and editing. The two experts operate over the same interleaved multimodal context, preserving cross-task interaction while avoiding direct competition between heterogeneous objectives. Task-specific heads are further used for autoregressive language modeling and flow-based visual generation, respectively.

Unified Context Learning. Lance first converts heterogeneous inputs into a shared interleaved multimodal sequence. (1) Text instructions are embedded using the language embedding layer of Qwen2.5-VL [5]. (2) For understanding-oriented visual inputs, Lance employs the Qwen2.5-VL ViT encoder [5], which uses $14\times$ spatial and $2\times$ temporal patching followed by a 2×2 spatial merge to produce compact semantic visual tokens. These tokens provide language-aligned visual semantics for multimodal understanding and reasoning. (3) For generation-oriented visual inputs, we encode images or videos into continuous latent representations using the Wan2.2 3D causal VAE encoder [109]. This encoder jointly supports image and video modalities through a unified latent space with $16\times$ spatial downsampling and $4\times$ temporal downsampling for videos. The resulting latent features preserve the low-level appearance and temporal structure required for high-fidelity visual generation, and are projected into the hidden space of the generation backbone through a lightweight MLP connector.

As a result, Lance represents each sample as a unified interleaved multimodal sequence of text tokens, ViT semantic tokens, clean VAE latent tokens, and noisy VAE latent tokens:

$$\mathcal{S} = \cdots \oplus \mathcal{B}_{\text{text}}(\mathbf{T}) \oplus \mathcal{B}_{\text{vis}}(\mathbf{V}_{\text{vit}}) \oplus \mathcal{B}_{\text{vis}}(\mathbf{V}_{\text{vae}}^{\text{clean}}) \oplus \mathcal{B}_{\text{vis}}(\mathbf{V}_{\text{vae}}^{\text{noisy}}) \oplus \mathcal{B}_{\text{text}}(\mathbf{T}') \oplus \cdots, \tag{1}$$

$$\mathcal{B}_{\text{text}}(\mathbf{T}) = [\texttt{BOT}, \mathbf{T}, \texttt{EOT}], \quad \mathcal{B}_{\text{vis}}(\mathbf{V}) = [\texttt{BOV}, \mathbf{V}, \texttt{EOV}]. \tag{2}$$

This formulation supports understanding, generation, and mixed interleaved multimodal samples within a single context modeling framework.

To handle such heterogeneous sequences, Lance adopts *generalized 3D causal attention*. The sequence is partitioned into modality-specific segments, where each segment attends to preceding clean segments to preserve causal dependencies. Within each segment, text tokens use causal attention, while visual tokens use bidirectional attention to capture spatial and spatiotemporal structure. This provides a unified attention mechanism for multimodal understanding, generation, and conditional editing.

Decoupled Capability Pathways. Although Lance organizes all modalities within a shared sequence, it processes understanding and generation through specialized expert pathways. The understanding expert LLM_{UND} primarily operates on text tokens and semantic visual tokens, and autoregressively predicts target text tokens for multimodal understanding. Its hidden states are mapped by a language modeling head and optimized with the standard next-token prediction loss:

$$\mathcal{L}_{\text{UND}} = -\sum_i \log p_{\theta_{\text{UND}}}(y_i \mid y_{<i}, \mathcal{S}). \tag{3}$$

The generation expert LLM_{GEN} operates on VAE latent tokens and predicts generation-side hidden states conditioned on the interleaved multimodal context. These hidden states are projected through an LLM-to-VAE connector into the latent space and passed to a flow prediction head. Let x_1 denote the clean VAE latent and $x_0 \sim \mathcal{N}(0, I)$ denote Gaussian noise. We construct the interpolated latent as $x_t = tx_1 + (1-t)x_0$, where $t \sim \mathcal{U}(0, 1)$, and optimize the generation expert with:

$$\mathcal{L}_{\text{GEN}} = \mathbb{E}_{x_0, x_1, t}\left[\left\| v_{\theta_{\text{GEN}}}(x_t, \mathcal{S}, t) - (x_1 - x_0) \right\|_2^2\right]. \tag{4}$$

Here, θ_{UND} and θ_{GEN} denote the pathway-specific parameters for understanding and generation, respectively, including their Transformer-decoder expert backbones and corresponding prediction heads.

The overall objective is:

$$\mathcal{L} = \lambda_u \mathcal{L}_{\text{UND}} + \lambda_g \mathcal{L}_{\text{GEN}}. \tag{5}$$

This design enables Lance to preserve unified context interaction while allowing semantic understanding and visual synthesis to specialize in their own representations, parameters, and objectives.

3.3 Modality-Aware Rotary Positional Encoding

Unified multimodal training places heterogeneous visual token groups within the same interleaved sequence, including ViT semantic tokens, clean VAE condition tokens, and noisy VAE target tokens. These tokens differ not only in their source encoders, but also in their functional roles: semantic tokens provide language-aligned visual cues for understanding, clean VAE latents serve as visual conditions, and noisy VAE latents are optimized as generation targets. Standard 3D-RoPE can encode spatiotemporal layouts, but it does not explicitly distinguish these heterogeneous token groups, which may lead to positional ambiguity and weaken cross-task alignment.

In the original 3D-RoPE formulation of Qwen2.5-VL [5], text tokens and visual tokens are assigned positional indices in different forms. Given N text tokens, the i-th text token is assigned $\mathbf{p}_i^{\text{text}} = [i, i, i]$. For visual tokens with temporal length T, height H, and width W, a token at location (t, h, w) is assigned a 3D position according to its spatiotemporal layout:

$$\hat{\mathbf{p}}_{t,h,w}^{\text{vis}} = N + [t,\ h,\ w] = [N + t,\ N + h,\ N + w], \tag{6}$$

where $t \in [0, T-1]$, $h \in [0, H-1]$, and $w \in [0, W-1]$.

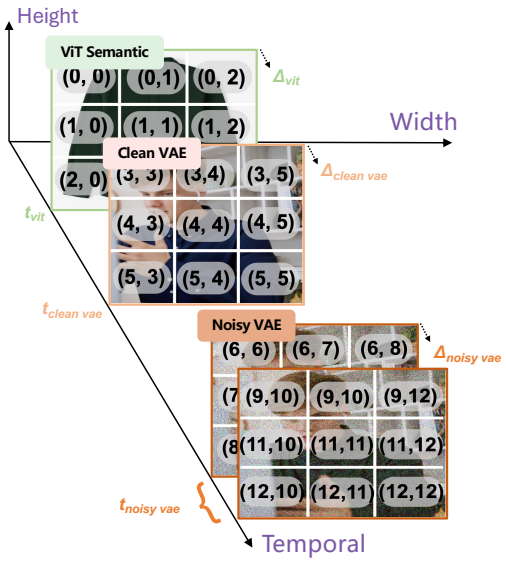


Figure 7 Illustration of modality-aware rotary positional encoding (MaPE).

This design is effective for standard image/video-language modeling. However, in unified multimodal training, a single sequence may contain multiple visual token groups from different modalities $\mathcal{M} = \{\mathbf{V}_{\text{vit}}, \mathbf{V}_{\text{vae}}^{\text{clean}}, \mathbf{V}_{\text{vae}}^{\text{noisy}}\}$. Assigning them only according to their spatiotemporal layouts may make their functional boundaries ambiguous in the positional space.

To address this issue, we introduce **Modality-Aware Rotary Positional Encoding** (MaPE), which injects token-group awareness into the positional indices. As shown in Figure 7, for each modality group $m \in \mathcal{M}$, we first define its base 3D-RoPE as $\hat{\mathbf{p}}_{t,h,w}^{(m)} = [\hat{t}_{t,h,w}^{(m)},\ \hat{h}_{t,h,w}^{(m)},\ \hat{w}_{t,h,w}^{(m)}]$, where the base coordinates follow the standard spatiotemporal assignment. MaPE then applies a modality-specific offset Δ_m only along the temporal dimension:

$$\mathbf{p}_{t,h,w}^{(m)} = \hat{\mathbf{p}}_{t,h,w}^{(m)} + [\Delta_m, 0, 0] = [\hat{t}_{t,h,w}^{(m)} + \Delta_m,\ \hat{h}_{t,h,w}^{(m)},\ \hat{w}_{t,h,w}^{(m)}]. \tag{7}$$

Applying modality offsets only to the temporal dimension provides two advantages. First, it explicitly separates different visual token groups in the global positional space, enabling the model to better distinguish the roles of semantic ViT features, clean VAE conditions, and noisy VAE targets. Second, since the spatial coordinates remain unchanged, the intrinsic spatial layouts within images and videos are preserved. Moreover, introducing modality offsets Δ_m along the t-dimension does not disrupt the temporal structure within a video. Since the offset is a shared constant shift for all tokens within the same modality group, the temporal order and relative distances of video latents are fully preserved. As a result, the model can better discriminate heterogeneous visual tokens while maintaining spatial consistency and temporal coherence.

	PT	CT	SFT	RL
Hyperparameters				
Learning rate	1.0×10^{-4}	1.0×10^{-4}	2.5×10^{-5}	2.0×10^{-6}
LR scheduler	Constant	Constant	Cosine	Constant
Weight decay	0.0	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.95,\ \epsilon = 1.0 \times 10^{-15}$)			
Loss weight (CE : MSE)	0.25 : 1	0.5 : 1	0.25 : 1	–
Warm-up steps	2500	2500	500	50
Training steps	350k	80k	15k	800
Sequence length per rank (min, max)	(44K, 50K)	(74K, 80K)	(74K, 80K)	(74K, 80K)
# Seen training tokens	1.5T	300B	72B	0.5B
Max context window	40k	70k	70k	70k
Gen resolution (min short side, max long side)	(192, 848)	(480, 848)	(480, 848)	(480, 848)
Und resolution (min short side, max long side)	(168, 826)	(462, 826)	(462, 826)	(462, 826)
Diffusion timestep shift	1.0	4.0	4.0	4.0

Table 2 Training hyperparameters of Lance.

Mixture	Ratio Type	PT	CT-I	CT-II	CT-III	SFT
Global	Vid.-Gen. : Vid.-Und. : Img.-Gen. : Img.-Und.	64 : 16 : 16 : 4	64 : 16 : 16 : 4	64 : 16 : 16 : 4	64 : 16 : 16 : 4	64 : 16 : 16 : 4
Generation	T2I : I-Edit : S2I	100 : 0 : 0	70 : 15 : 15	60 : 20 : 20	50 : 25 : 25	60 : 20 : 20
	T2V : I2V : V-Edit : S2V	100 : 0 : 0 : 0	60 : 10 : 15 : 15	40 : 20 : 20 : 20	25 : 25 : 25 : 25	60 : 10 : 15 : 15

Table 3 Training data mixture schedule of Lance. Img., Vid., Gen., and Und. denote image, video, generation, and understanding, respectively. CT is divided into three stages that progressively increase the proportion of challenging generation tasks.

4 Training and Data

Lance adopts a staged multi-task training strategy to progressively develop and balance multimodal understanding and generation within a unified task formulation. As shown in Table 2, the pipeline consists of four stages: PT establishes basic image/video understanding and generation from large-scale paired data; CT expands the task space with interleaved multi-task data and promotes cross-task transfer; SFT refines instruction following, visual fidelity, editing accuracy, and identity consistency with curated supervision; and RL further optimizes image generation with task-specific rewards. The data mixture schedule and task statistics are summarized in Tables 3 and 4.

4.1 Pre-Training Stage (PT)

Training Objectives. The pre-training stage establishes preliminary multimodal alignment and basic visual generation capabilities. To this end, we freeze the VAE and ViT encoders and optimize the remaining components, including the multimodal backbone, QK-Norm modules, and MLP connectors.

Pre-Training Data. The PT stage is trained on large-scale image-text and video-text pairs, organized around paired captioning and conditional generation tasks. The image-text subset comprises approximately 1B samples spanning diverse visual domains, including natural scenes, human-centric, object-centric, knowledge-oriented, and stylized content. The video-text subset comprises approximately 140M samples and covers diverse dynamic scenarios, including actions, events, scene transitions, and long-range temporal processes. To improve scalability, we adopt a progressive resolution curriculum of 192p $\rightarrow$ 360p $\rightarrow$ 480p, with dynamic resolution enabled at each stage. In addition, we use an image:video sampling ratio of approximately 1 : 4 to account for the greater difficulty of video modeling and to strengthen temporal reasoning and generation.

Output Type	Notation	Task	# Samples	Phases
Text	I2T	General image captioning	1B	PT, CT
	V2T	General video captioning	140M	PT, CT
	I2T	High-quality image captioning	190K	SFT
	V2T	High-quality video captioning	5K	SFT
	X2T	Interleaved multimodal understanding	2.73M	CT, SFT
Image	T2I	General image generation	1B	PT, CT
	X2I	General image editing	2.8M	CT
	X2I	General subject-driven image generation	3.6M	CT
	T2I	High-quality image generation	190K	SFT
	X2I	High-quality image editing	84K	SFT
Video	T2V/I2V	General video generation	140M	PT, CT
	X2V	General video editing	2.6M	CT
	X2V	General subject-driven video generation	1M	CT
	T2V/I2V	High-quality video generation	5K	SFT
	X2V	High-quality video editing	9K	SFT
	X2V	High-quality subject-driven video generation	5.5K	SFT

Table 4 Summary of task categories and sample statistics for Lance. Within each output type, high-quality data are listed separately and highlighted in gray. "Phases" indicates the training phase(s) where each data type is applied.

System Prompt for I2T/V2T captioning tasks

```
<|im_start|>system
Generate a detailed and accurate description of the {image/video}, including all the visual
details {and key moments}.<|im_end|>
<|im_start|>user
<|vision_start|><|user_vision|><|vision_end|><|im_end|>
<|im_start|>assistant
```

System Prompt for other I2T/V2T tasks

```
<|im_start|>system
View the {image/video} attentively and provide a suitable answer to the posed
question.<|im_end|>
<|im_start|>user
<|vision_start|><|user_vision|><|vision_end|><|user_text|><|im_end|>
<|im_start|>assistant
```

Figure 8 System prompts for understanding tasks. Red placeholders denote user-provided text and visual inputs.

4.2 Continual Training Stage (CT)

Training Objectives. The continual training stage extends the PT model from basic paired supervision to unified multi-task multimodal learning. By introducing richer interleaved multimodal data and more diverse input-output mappings, CT expands the task space and improves task-aware multimodal generalization.

Continual Training Data. During CT, we progressively introduce a broader set of tasks for both understanding and generation. For understanding, we incorporate 2.73M interleaved multimodal understanding samples, covering pure text understanding (T2T, 41K), captioning (443K), classification (142K), conversation (72K),

System Prompt for T2I/T2V tasks

```
<|im_start|>system
Describe the {image/video} by detailing the color, quantity, text, shape, size, texture,
spatial relationships {and motion/camera movements} of the objects and background:<|im_end|>
<|im_start|>user
<|user_text|><|im_end|>
<|im_start|>assistant
```

System Prompt for other X2I/X2V tasks

```
<|im_start|>system
Describe the key features of the input {image/video} (color, shape, size, texture, objects,
background), then explain how the user's text instruction should alter or modify the
{image/video}.  Generate a new {image/video} that meets the user's requirements while
maintaining consistency with the original input where appropriate.<|im_end|>
<|im_start|>user
<|vision_start|><|user_vision|><|vision_end|><|user_text|><|im_end|>
<|im_start|>assistant
```

Figure 9 System prompts for generation tasks. Red placeholders denote user-provided text and visual inputs.

grounding (200K), reasoning (194K), VQA (600K), and OCR (120K). For generation, we incorporate large-scale any-to-image/video data, including 2.8M image editing samples and 2.6M video editing samples, together with 3.6M subject-driven image generation samples and 1M subject-driven video generation samples. To accommodate the increased task diversity, we adopt a progressive data-mixture strategy that gradually increases the sampling ratio of more challenging tasks, such as editing and subject-driven generation, while correspondingly reducing the proportion of simpler caption-style supervision (detailed in Table 3). In total, the CT stage consumes approximately 300B training tokens.

Task-specific System Prompts. To better distinguish heterogeneous tasks within a unified multimodal context, we further introduce task-specific *system prompts* for understanding and generation tasks, as illustrated in Figure 8 and Figure 9. These prompts provide explicit task priors and guide task-specific input-output formats while preserving unified sequence modeling.

4.3 Supervised Fine-Tuning Stage (SFT)

Training Objectives. The supervised fine-tuning stage refines the model with high-quality, task-aligned supervision under a reduced learning rate. Unlike PT and CT, which focus on capability acquisition and task expansion, SFT emphasizes instruction fidelity, visual consistency, editing accuracy, and identity preservation, improving controllability and downstream task performance.

Supervised Fine-Tuning Data. The SFT stage uses curated high-quality data spanning both understanding and generation tasks. For understanding, we use 190K high-quality image captioning samples, 5K high-quality video captioning samples, together with 2.73M interleaved multimodal understanding samples for continued instruction refinement. For image generation, we include 190K high-quality image generation samples and 84K high-quality image editing samples. For video generation, we further incorporate 5K high-quality video generation samples, 9K high-quality video editing samples, and 5.5K high-quality subject-driven video generation samples. Compared with the large-scale corpora used in PT and CT, these curated data provide stronger task alignment and higher annotation quality, and thus offer more precise supervision for improving instruction following and generation fidelity.

Models	Params.	DPG-Bench						GenEval						
		Global	Entity	Attribute	Relation	Other	Overall	1-Obj.	2-Obj.	Count	Colors	Position	Attr.	Overall
Generation-only Models														
PixArt-α [12]	0.6B	74.97	79.32	78.60	82.57	76.96	71.11	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SDXL [91]	3.5B	83.27	82.43	80.91	86.76	80.41	74.65	0.98	0.74	0.39	0.85	0.15	0.23	0.55
Hunyuan-DiT [62]	1.5B	84.59	80.59	88.01	74.36	86.41	78.87	–	–	–	–	–	–	–
DALL-E 3 [6]	–	90.97	89.61	88.39	90.58	89.83	83.50	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium [25]	2B	87.90	91.01	88.83	80.70	88.68	84.08	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Emu3-Gen [115]	8B	85.21	86.68	86.84	90.22	83.15	80.60	0.98	0.71	0.34	0.81	0.17	0.21	0.54
FLUX.1-dev† [53]	12B	74.35	90.00	88.96	90.87	88.33	83.84	0.98	0.93	0.75	0.93	0.68	0.65	0.82
GPT Image 1 [87]	–	–	–	–	–	–	–	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Qwen-Image [122]	20B	91.32	91.56	92.02	94.31	92.73	88.32	0.99	0.92	0.89	0.88	0.76	0.77	0.87
Unified Models														
SEED-X [32]	–	–	–	–	–	–	–	0.97	0.58	0.26	0.80	0.19	0.14	0.49
TokenFlow-XL [92]	–	–	–	–	–	–	–	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus [123]	–	82.33	87.38	87.70	85.46	86.41	79.68	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Emu3-Gen† [115]	8B	–	–	–	–	–	81.60	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Show-o [133]	–	–	–	–	–	–	–	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Janus-Pro-7B [14]	7B	86.90	88.90	89.40	89.32	89.48	84.19	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Ovis-U1 [111]	1.2B	82.37	90.08	88.68	93.35	85.20	83.72	–	–	–	–	–	–	–
OmniGen2 [124]	4B	88.81	88.83	90.18	89.37	90.27	83.57	**1.00**	0.95	0.64	0.88	0.55	0.76	0.80
Show-o2 [134]	7B	89.00	**91.78**	89.96	91.81	**91.64**	86.14	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76
UniWorld-V1 [65]	13B	83.64	88.39	88.44	89.27	87.22	81.38	0.99	0.93	0.79	0.89	0.49	0.70	0.80
BAGEL† [22]	7B	88.94	90.37	91.29	90.82	88.67	85.07	0.98	0.95	**0.84**	0.95	0.78	0.77	0.88
Mogao [63]	7B	82.37	90.03	88.26	93.18	85.40	84.33	**1.00**	**0.97**	0.83	0.93	0.84	0.80	0.89
InternVL-U [103]	1.7B	90.39	90.78	90.68	90.29	88.77	85.18	0.99	0.94	0.74	0.91	0.77	0.74	0.85
TUNA [77]	7B	**90.42**	91.68	90.94	91.87	90.73	**86.76**	**1.00**	**0.97**	0.81	0.91	**0.88**	**0.83**	**0.90**
TUNA-2 [78]	7B	89.50	91.40	**92.07**	91.91	88.81	86.54	0.99	0.96	0.80	0.91	0.84	0.76	0.87
Lance (Ours)	**3B**	83.89	91.07	89.36	**93.38**	80.80	84.67	**1.00**	0.94	**0.84**	**0.97**	0.87	0.81	**0.90**

Table 5 Image generation results on DPG-Bench and GenEval. † refers to methods using LLM rewriters in GenEval. **Bold**: best results among unified models. Underline: second-best among unified models.

4.4 Reinforcement Learning Stage

Training Objectives. The reinforcement learning stage further refines the model's image generation capability by directly optimizing generation behavior with task-specific rewards. Unlike SFT, which learns from static supervised targets through maximum likelihood, RL uses Group Relative Policy Optimization (GRPO) to encourage outputs that better satisfy fine-grained textual constraints. In particular, this stage focuses on improving text rendering accuracy, image-text correspondence, and prompt compositional adherence.

Reinforcement Learning Data. The RL stage uses 20K image generation prompts that emphasize fine-grained text-related requirements. During optimization, PaddleOCR [18] serves as the reward model to evaluate the consistency between the generated image and the textual constraints specified in the prompt. This reward provides direct feedback on text rendering quality and text-image alignment, helping improve aspects that are difficult to fully capture with supervised fine-tuning alone.

5 Experiments

5.1 Experimental Setup

Lance is implemented upon Qwen2.5-VL 3B [5], using its weights to initialize the visual understanding encoder and the multimodal context backbones LLM_{UND} and LLM_{GEN}. For the visual generation encoder, we adopt

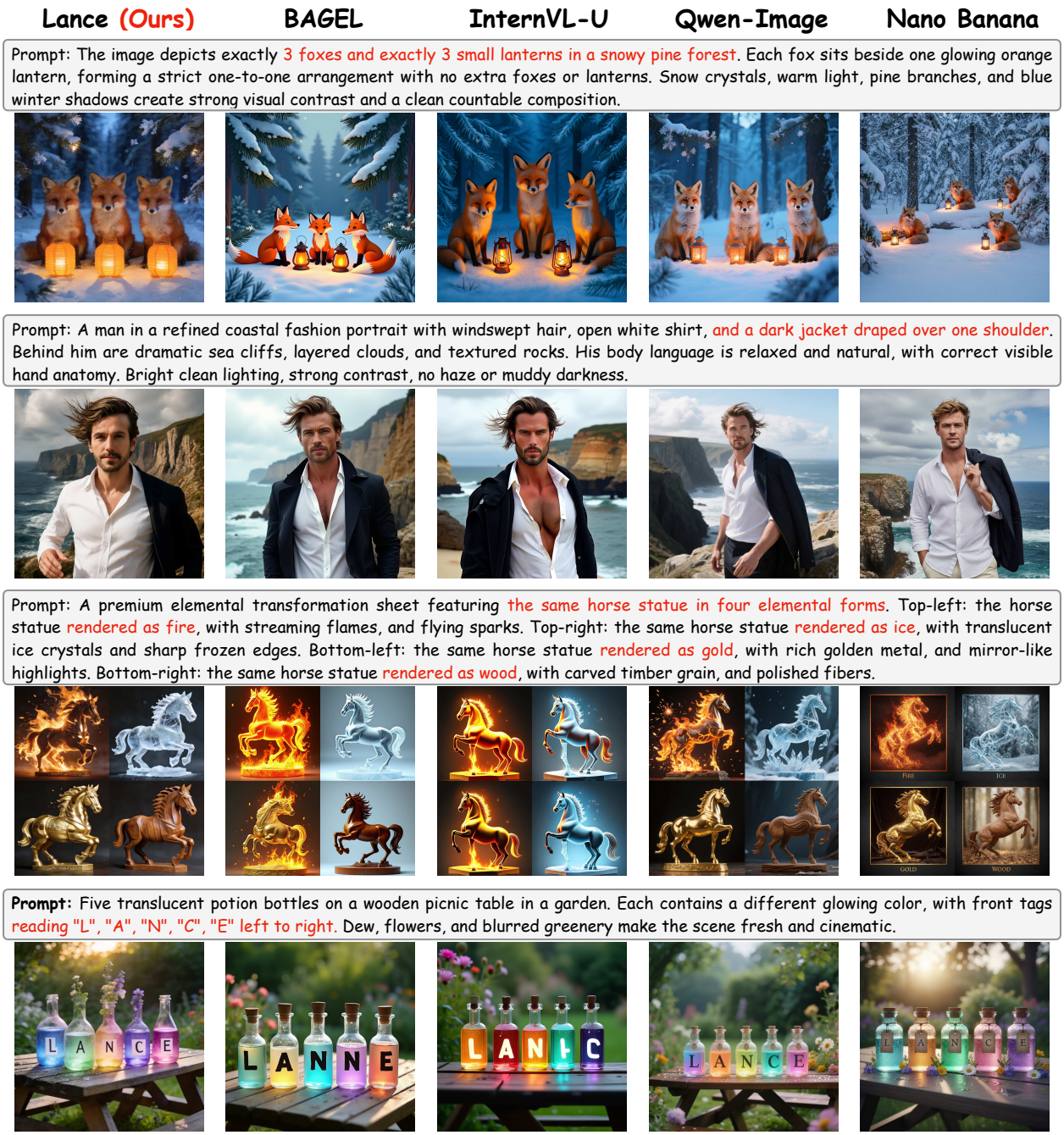


Figure 10 T2I qualitative comparison. Instructions that are correctly reflected in our results but missed or incorrectly rendered by some baseline models are highlighted in **red**.

the 3D causal VAE encoder from Wan2.2 [109], to support a unified processing of image and video modalities. Following prior work [38], we also adopt classifier-free guidance (CFG) for visual and text conditions. During the PT stage, for text-to-image generation data, the text condition is dropped with a probability of 10%. During the CT and SFT stages, for multimodal conditions, the full condition is dropped with a probability of 5%, while the text-only condition is additionally dropped with a probability of 5% and the visual condition is retained. During inference, the CFG scale for text conditions in generation tasks is set to 4. Unless otherwise specified, the image input resolution is set to 768×768, while videos are sampled at $480p$ resolution with a

frame rate of 12 fps.

5.2 Main Results

5.2.1 Image Generation

Quantitative Results. We evaluate the image generation capability of Lance on GenEval [33] and DPG-Bench [41]. As shown in Table 5, Lance achieves top-tier performance among unified models on GenEval, matching the best overall score (**0.90**) while showing strong compositional ability on counting, colors, and spatial position. On DPG-Bench, Lance obtains competitive overall performance and performs particularly well on relation modeling, indicating its ability to preserve fine-grained semantic consistency under complex prompts. These results suggest that Lance can effectively support high-quality image synthesis within a unified multimodal framework, despite using only 3B activated parameters.

Qualitative Results. We conduct a qualitative comparison of Lance with 7B Bagel [22], 1.7B InternVL-U [103], 20B Qwen-Image [122] and Nano Banana [34]. As shown in Figure 10, compared with open-source unified multimodal baselines such as Bagel [22] and InternVL-U [103], Lance demonstrates stronger visual aesthetics and image-text alignment (*e.g.*, lantern count in 1-st case, jacket draped over one shoulder in 2-nd case). Overall, Lance generates significantly higher-quality images than Bagel [22] and InternVL-U [103], and achieves comparable performance with the 20B large-scale model Qwen-Image [122] and the commercial closed-source model Nano Banana [34].

5.2.2 Video Generation

Quantitative Results. We evaluate the text-to-video generation capability of Lance on VBench [46]. As shown in Table 6, Lance achieves the best Total Score (**85.11**) among unified models with only 3B activated parameters. Beyond the overall score, Lance also shows strong performance across both quality-oriented and semantic-oriented dimensions, including visual quality, object grounding, color consistency, spatial relationships, scene understanding, and temporal style. These results indicate that the proposed unified framework effectively supports compositional video generation and text-video alignment, while scaling naturally from image generation to more challenging spatiotemporal generation tasks.

Qualitative Results. We conduct a qualitative comparison between Lance and 8.3B HunyuanVideo1.5 [121], 5B Wan2.2-TI2V [109], and 7B UniVideo [119]. As shown in Figure 11, the generated videos exhibit strong semantic fidelity, coherent motion, and appealing visual quality. In challenging cases involving complex human interactions (*e.g.*, 1-st case, “two adults hugging"), or explicit camera transitions (*e.g.*, 2-nd case, from a “medium view" to “close facial framing"), our model follows the prompt accurately and produces videos with stable visual texture and consistent temporal evolution. These examples further demonstrate the effectiveness of the unified architecture for high-quality text-to-video generation.

5.2.3 Multimodal Editing

Quantitative Results. We evaluate the image editing capability of our model on GEdit-Bench [76]. As shown in Table 7, our model achieves the best Avg/G_O score (7.30) among unified models, demonstrating strong overall editing performance under a compact parameter budget. In particular, our model obtains the best results in several key editing categories, including background change, material modification, motion change, portrait beautification, subject removal, replacement, and tone transfer. These results suggest that the proposed unified framework can effectively support a broad range of image editing operations. We also observe that Lance is relatively weaker on text modification, indicating that text-specific editing remains an important direction for future improvement.

Qualitative Results. We further provide qualitative results for both image and video editing in Figure 12. For image editing, Lance achieves visually coherent image editing with well-preserved structures and realistic textures, *e.g.*, the plausible hand geometry and fine details in the 2-nd case. For video editing, Lance performs accurate multi-attribute modifications while maintaining natural motion dynamics, such as the temporally consistent hand movement of the person holding a cup in the last case. Overall, these results demonstrate

(a) VBench Metrics Part I

Models	Params.	Quality Score	Semantic Score	Subj. Consist.	Bkg. Consist.	Temp. Flicker.	Motion Smooth.	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
Generation-only Models											
ModelScope [112]	1.7B	78.05	66.54	89.87	95.29	98.28	95.79	66.39	52.06	58.57	82.25
LaVie [116]	3B	78.78	70.31	91.41	97.47	98.30	96.38	49.72	54.94	61.90	91.82
Show-1 [143]	6B	80.42	72.98	95.53	98.02	99.12	98.24	44.44	57.35	58.66	93.07
AnimateDiff-V2 [35]	–	82.90	69.75	95.30	97.68	98.75	97.76	40.83	67.16	70.10	90.90
VideoCrafter-2.0 [10]	–	82.20	73.42	96.85	98.22	98.41	97.73	42.50	63.13	67.22	92.55
CogVideoX [140]	5B	82.75	77.04	96.23	96.52	98.66	96.92	70.97	61.98	62.90	85.23
Kling [50]	–	83.39	75.68	98.33	97.60	99.30	99.40	46.94	61.21	65.62	87.24
Open-Sora-2.0 [90]	–	82.10	80.14	98.75	98.00	99.40	99.49	20.74	64.33	65.62	94.50
Gen-3 [96]	–	84.11	75.17	97.10	96.62	98.61	99.23	60.14	63.34	66.82	87.81
Step-Video-T2V [79]	30B	84.46	71.28	98.05	97.67	99.40	99.08	53.06	61.23	70.63	80.56
HunyuanVideo [121]	–	85.07	76.88	97.22	97.60	99.39	99.05	71.94	60.28	67.24	83.48
Wan2.1-T2V [109]	14B	85.59	76.11	97.52	98.09	99.46	98.30	65.46	66.07	69.43	86.28
Unified Models											
HaploOmni [132]	7B	–	–	96.40	97.60	–	96.80	65.30	–	–	–
Emu3 [115]	8B	–	–	95.32	**97.69**	–	**98.93**	**79.27**	59.64	–	86.17
VILA-U [130]	7B	76.26	65.04	–	–	–	–	–	–	–	–
Show-o2 [134]	2B	82.10	78.31	**97.28**	96.78	97.68	98.25	40.83	65.15	**67.06**	94.81
TUNA [77]	1.5B	84.32	83.04	95.99	96.72	98.02	98.33	69.39	**65.88**	66.83	95.41
Lance (Ours)	**3B**	**85.14**	**84.96**	94.52	94.28	**99.66**	95.93	75.83	64.33	66.78	**96.58**

(b) VBench Metrics Part II

Models	Params.	Multi. Objects	Human Action	Color	Spatial Relation	Scene	Appear. Style	Temp. Style	Overall Consist.	Total Score↑
Generation-only Models										
ModelScope [112]	1.7B	38.98	92.40	81.72	33.68	39.26	23.39	25.37	25.67	75.75
LaVie [116]	3B	33.32	96.80	86.39	34.09	52.69	23.56	25.93	26.41	77.08
Show-1 [143]	6B	45.47	95.60	86.35	53.50	47.03	23.06	25.28	27.46	78.93
AnimateDiff-V2 [35]	–	36.88	92.60	87.47	34.60	50.19	22.42	26.03	27.04	80.27
VideoCrafter-2.0 [10]	–	40.66	95.00	92.92	35.86	55.29	25.13	25.84	28.23	80.44
CogVideoX [140]	5B	62.11	99.40	82.81	66.35	53.20	24.91	25.38	27.59	81.61
Kling [50]	–	68.05	93.40	89.90	73.03	50.86	19.62	24.17	26.42	81.85
Open-Sora-2.0 [90]	–	77.72	95.40	85.98	76.18	52.71	22.98	25.91	27.57	81.71
Gen-3 [96]	–	53.64	96.40	80.90	65.09	54.57	24.31	24.71	26.69	82.32
Step-Video-T2V [79]	30B	50.55	94.00	88.25	71.47	24.38	23.17	26.01	27.12	81.83
HunyuanVideo [121]	–	66.71	94.40	89.79	72.13	54.46	22.21	24.52	26.95	83.43
Wan2.1-T2V [109]	14B	69.58	95.40	88.59	75.39	45.75	22.64	23.19	25.91	83.69
Unified Models										
HaploOmni [132]	7B	–	–	–	–	34.60	–	–	–	78.10
Emu3 [115]	8B	44.64	77.71	–	68.73	37.11	20.92	–	–	80.96
VILA-U [130]	7B	–	–	–	–	–	–	–	–	74.01
Show-o2 [134]	2B	76.01	95.20	80.89	62.61	57.67	**23.29**	25.27	27.00	81.34
TUNA [77]	1.5B	92.31	97.50	87.67	78.12	58.59	23.18	24.68	**27.71**	84.06
Lance (Ours)†	**3B**	**93.86**	**97.80**	**92.61**	**93.61**	**64.75**	23.14	**25.53**	27.04	**85.11**

Table 6 Video generation results on VBench. † refers to methods using LLM rewriters. **Bold**: best results among unified models. Underline: second-best among unified models.

Figure 11 T2V qualitative comparison. Instructions that are correctly reflected in our results but missed or incorrectly rendered by some baseline models are highlighted in **red**.

Models	Params.	GEdit-Bench											
		BC	CA	MM	MC	PB	ST	SA	SR	SRp	TM	TT	Avg/G_O
Generation-only Models													
Gemini 2.0 [102]	–	–	–	–	–	–	–	–	–	–	–	–	6.32
GPT Image 1 [87]	–	6.96	6.85	7.10	5.41	6.74	7.44	7.51	8.73	8.55	8.45	8.69	7.49
Qwen-Image-Edit [122]	20B	8.23	8.30	7.33	8.05	7.49	6.74	8.57	8.09	8.29	8.48	8.50	8.01
Unified Models													
Lumina-DiMOO [135]	8B	3.43	4.27	3.08	2.77	4.74	5.19	4.44	3.80	4.38	2.68	4.20	3.91
Ovis-U1 [111]	1.2B	7.49	6.88	6.21	4.79	5.98	6.46	7.49	7.25	7.27	4.48	6.31	6.42
BAGEL [22]	7B	7.32	6.91	6.38	4.75	4.57	6.15	**7.90**	7.16	7.02	7.32	6.22	6.52
InternVL-U [103]	1.7B	7.08	7.05	6.38	7.02	6.03	6.27	7.13	6.55	6.33	6.59	6.85	6.66
InternVL-U (w/ CoT) [103]	1.7B	7.05	**7.87**	6.50	6.99	5.77	6.10	7.33	7.16	7.12	**7.36**	6.46	6.88
Lance (Ours)	**3B**	**7.73**	7.74	**7.28**	**7.83**	**7.50**	**7.03**	7.64	**7.85**	**7.71**	4.46	**7.57**	**7.30**

Table 7 Image editing results on GEdit-Bench. Bold: best results among unified models. Underline: second-best among unified models.

Models	Params.	MVBench																			
		AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	CO	EN	ER	CI	Avg.↑
Understanding-only Models																					
Video-LLaMA [144]	7B	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	40.0	30.0	21.0	37.0	34.1
LLaMA-Adapter [145]	7B	23.0	28.0	51.0	30.0	33.0	53.5	32.5	33.5	25.5	21.5	30.5	29.0	22.5	41.5	39.5	31.5	22.5	28.0	32.0	31.7
Video-ChatGPT [81]	7B	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	33.0	29.5	26.0	35.5	32.7
VideoChat [60]	7B	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	41.0	23.5	23.5	36.0	35.5
VideoChat2 [59]	7B	66.0	47.5	83.5	49.5	60.0	58.0	71.5	42.5	23.0	23.0	88.5	39.0	42.0	58.5	44.0	36.5	35.0	40.5	65.5	51.1
ST-LLM [75]	7B	66.0	53.5	84.0	44.0	58.5	80.5	73.5	38.5	42.5	31.0	86.5	36.5	56.5	78.5	43.0	46.5	34.5	41.5	58.5	54.9
GPT-4V [86]	–	55.5	63.5	72.0	46.5	73.5	18.5	59.0	29.5	12.0	40.5	83.5	39.0	12.0	22.5	45.0	52.0	31.0	59.0	11.0	43.5
PLLaVA [136]	34B	67.5	53.0	82.0	47.0	79.0	68.5	67.5	36.5	37.5	49.5	91.0	40.5	43.0	70.0	51.5	66.5	39.5	63.5	59.0	58.1
Video-CCAM [27]	9B	83.0	67.0	89.5	49.0	72.0	86.5	81.0	45.0	28.0	29.0	90.0	59.0	67.0	85.0	63.5	77.0	34.0	73.5	59.0	64.6
Qwen2.5-VL [5]	3B	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	67.0
TimeMarker [13]	8B	79.0	74.5	89.0	53.5	77.0	94.0	76.0	41.5	52.5	47.0	91.5	53.0	76.5	92.5	57.0	70.5	23.5	53.5	82.5	67.4
InternVideo2 [117]	7B	86.0	70.0	87.0	56.0	75.0	91.0	86.0	40.0	48.0	53.0	90.0	41.0	73.0	92.0	52.0	56.0	33.0	57.0	74.0	67.3
Unified Models																					
Show-o2 [134]	1.5B	63.8	59.5	63.5	40.0	70.5	54.5	66.0	36.5	36.0	27.0	88.0	43.5	43.0	58.0	44.5	54.0	28.5	39.5	45.0	50.6
Show-o2 [134]	7B	60.1	67.0	68.0	45.5	**78.0**	51.0	**73.5**	**44.5**	36.0	**39.0**	**92.5**	**51.5**	36.0	59.5	**52.0**	**64.0**	**38.0**	**60.0**	43.0	55.7
TUNA [77]	1.5B	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	54.4
UniVideo [119]	7B	54.3	41.5	**77.5**	**50.0**	62.5	68.2	50.5	37.5	36.0	29.5	35.5	28.5	52.5	70.5	33.5	40.5	37.5	36.5	38.0	46.3
Lance (Ours)	**3B**	**73.9**	**76.5**	71.5	49.0	63.5	**96.0**	72.5	33.0	**63.5**	33.0	86.0	41.0	**82.0**	**97.5**	43.0	47.5	31.5	40.0	**77.0**	**62.0**

Table 8 Video understanding results on MVBench. Bold: best results among unified models. Underline: second-best among unified models.

Lance's high-fidelity editing ability in both spatial realism and temporal coherence, highlighting the potential of unified models for multimodal editing.

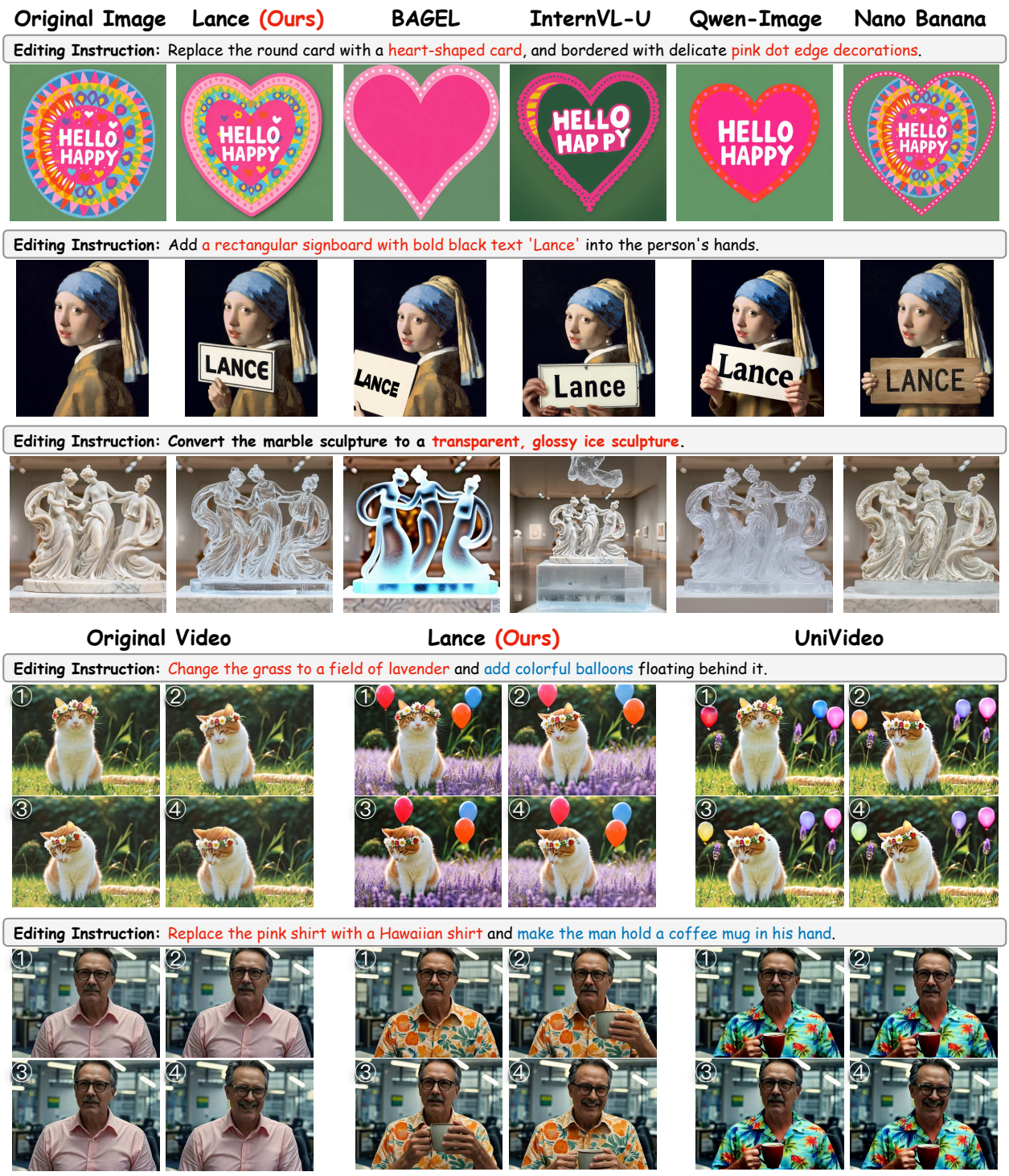


Figure 12 Multimodal editing qualitative comparison. Lance performs precise image editing with realistic texture and structure preservation, and supports temporally coherent video editing with natural motion dynamics.

5.2.4 Multimodal Understanding

Quantitative Results. We evaluate the video understanding ability of Lance on MVBench [59], a widely used multi-choice benchmark for assessing temporal perception and video-centric understanding. As reported in

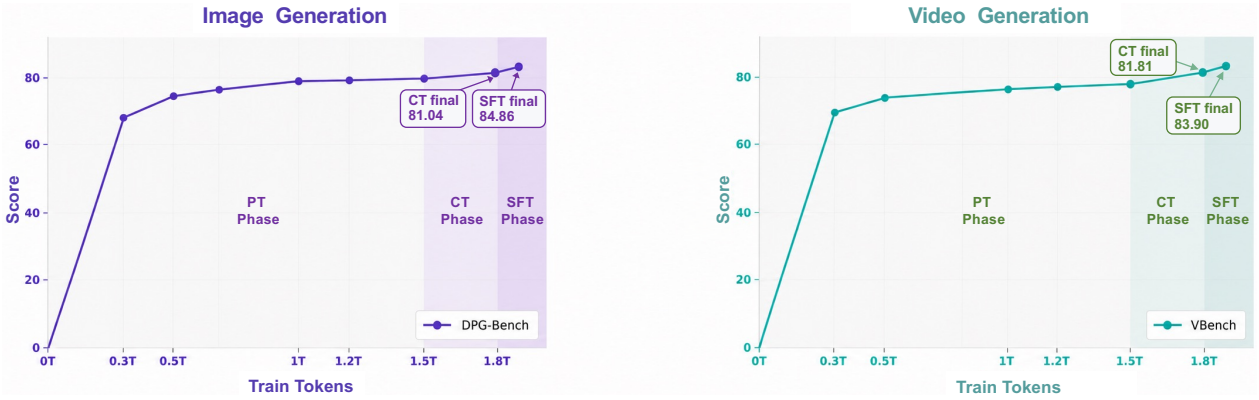


Figure 13 Scaling behavior of image and video generation performance with increasing training tokens. We report DPG-Bench for image generation and VBench for video generation across different training token budgets.

Ablation Type	Setting	Image Generation	Video Generation
		GenEval ↑	VBench ↑
Base	Gen. only	80.88	81.25
+ Understanding data	Gen.:Und. = 8:2	81.65	<u>82.91</u>
	Gen.:Und. = 9:1	80.93	81.47
+ Multi-task data	Gen.:MT-Gen. = 8:2	<u>81.89</u>	82.88
	Gen.:MT-Gen. = 6:4	**82.06**	**83.05**

Table 9 Ablation on cross-task data for generation. Gen. denotes base generation data, Und. denotes understanding data, and MT-Gen. denotes multi-task generation data, including editing, subject-driven generation, etc.

Table 8, Lance achieves the highest overall score (**62.0**) among existing unified multimodal models, with an approximately **11.3%** relative improvement compared to the second-best unified model, Show-o2 7B [134]. Lance also surpasses most of the specialized understanding models, with only half or even fewer parameters, indicating that unified multi-task training can preserve strong video understanding while enabling generation and editing capabilities.

Qualitative Results. We present qualitative examples for image and video understanding in Figures 3 and 5. Lance handles diverse understanding tasks, including OCR, knowledge-grounded reasoning, multi-image motion analysis, detailed video captioning, and action counting. The examples show that Lance can recognize fine-grained visual details, reason over static images, and capture temporal dynamics in videos. These results indicate that Lance maintains strong multimodal understanding ability while jointly supporting generation and editing within a unified model.

6 Ablation Study

6.1 Training Dynamics Analysis

To systematically analyze the evolution of model capabilities during training, we further conduct quantitative and qualitative evaluations of model variants under different training-token budgets.

Quantitative Analysis. As shown in Figure 13, image and video generation exhibit broadly consistent scaling trends as training tokens increase, with rapid gains in the early PT stage followed by a slower-growth regime. This indicates that large-scale paired training first establishes core generation capability, while later tokens mainly refine prompt alignment, visual fidelity, and temporal consistency. Moreover, the CT stage further improves native generation capability, even though it mainly introduces multi-task data such as editing and

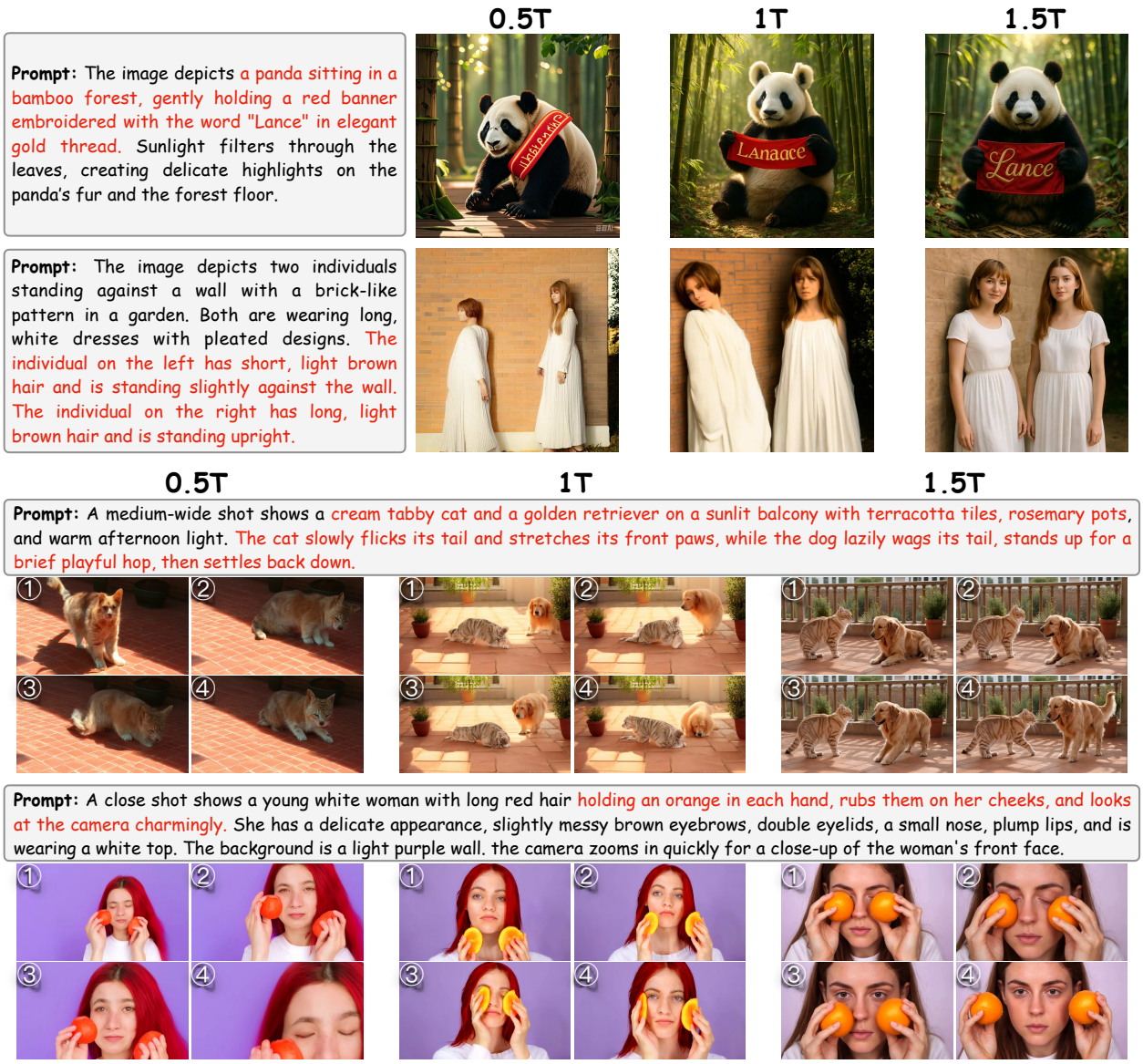


Figure 14 Comparison of model variants trained with different token budgets. We present qualitative cases of text-to-image and video generation using model variants trained with 0.5T, 1T, and 1.5T tokens. As the training budget increases, the model demonstrates improved prompt alignment, visual fidelity, and temporal consistency.

instruction-following data rather than additional pure generation data (Table 4). These results suggest that multi-task integration not only strengthens editing and instruction-following behaviors, but also brings positive transfer to visual generation, further validating the role of multi-task synergy in enhancing unified multimodal modeling.

Qualitative Analysis. Figure 14 shows visual results consistent with the quantitative trends. As the training budget increases from 0.5T to 1.5T, Lance progressively improves prompt alignment, visual fidelity, text rendering, and temporal coherence. Early models capture coarse semantics but still suffer from distorted text, inaccurate attributes, and unstable motion, while the 1.5T model produces more faithful compositions and more coherent multi-object dynamics.

Setting	Image Generation	Image Editing	Video Generation	Video Understanding
	GenEval ↑	GEdit ↑	VBench ↑	MVBench ↑
w/ MaPE	**80.94**	**6.86**	**81.81**	**59.16**
w/o MaPE	80.56	6.30	80.95	59.02

Table 10 Ablation on Modality-Aware Rotary Positional Encoding (MaPE). We report GenEval for image generation, GEdit for image editing, VBench for video generation, and MVBench for video understanding.

6.2 Effect of Cross-Task Data Synergy

We conduct ablation studies to further analyze how different task mixtures affect the generation ability of Lance, focusing on the effects of understanding data and multi-task generation data. The results are summarized in Table 9.

Effect of Understanding Data. Introducing understanding-oriented data brings clear gains when used at an appropriate ratio. In particular, the Gen.:Und. $= 8:2$ setting improves both image and video generation, suggesting that understanding data provides useful semantic grounding for visual synthesis.

Effect of Multi-task Data. Multi-task generation data enhances the base generation capability via joint training. Both mixture ratios outperform the generation-only baseline, with Gen.:MT-Gen. $= 6:4$ achieving the best overall results. This indicates that multi-task synergy provides complementary supervision that promotes more robust visual composition and fine-grained synthesis.

6.3 Effect of Modality-Aware Rotary Positional Encoding

We further ablate the proposed Modality-Aware Rotary Positional Encoding (MaPE) to verify its effectiveness in unified multimodal modeling. As shown in Table 10, removing MaPE consistently degrades performance across generation, editing, and understanding. The improvement is especially clear on image editing (from 6.30 to 6.86), where the model needs to jointly reason over visual conditions and generation targets. This suggests that MaPE reduces positional ambiguity among heterogeneous visual token groups, leading to better cross-task contextual alignment and more stable visual synthesis.

7 Conclusion, Limitations and Future Work

In this work, we present Lance, a lightweight native unified multimodal model for image and video understanding, generation, and editing. Our key finding is that multi-task synergy can effectively advance unified multimodal modeling, enabling diverse tasks to mutually enhance each other within a shared framework. To this end, Lance combines unified interleaved context modeling with decoupled capability pathways, allowing semantic understanding and visual synthesis to interact while preserving task-specific specialization. Extensive experiments demonstrate that Lance achieves strong performance across image generation, video generation, multimodal editing, and video understanding benchmarks. Notably, these results are obtained with only 3B activated parameters and a maximum 128-GPU training budget, showing that capable unified multimodal models can be built in a resource-efficient manner.

Lance opens several promising directions for future exploration.

- **Post-training:** More comprehensive video-aware reward models, together with reward-based optimization methods [74, 137, 148], could provide stronger supervision for temporally coherent, visually appealing, and user-aligned generation.
- **Model Scaling:** Scaling model capacity, expert capacity, and context length may further improve Lance's overall capability and cross-task generalization.
- **Broader Modalities:** Incorporating audio, speech, 3D, depth, and embodied sensory signals would be a natural step toward general-purpose any-to-any multimodal intelligence.

- **Streaming Multimodal Interaction:** Integrating streaming perception and generation mechanisms [45, 108, 120] could extend Lance toward real-time interaction and closed-loop multimodal agents.

We hope Lance can serve as a practical foundation for future research on efficient, scalable, and task-general unified multimodal systems.

Author Contributions. Fengyi Fu, Mengqi Huang, Shaojin Wu, Yufei Huo and Jianzhu Guo contributed to code development, algorithm design, model training, and evaluation. Jianzhu Guo and Mengqi Huang initialized the codebase. Fengyi Fu, Mengqi Huang, Jianzhu Guo and Shaojin Wu were involved in the pre-training, continued training, and supervised fine-tuning stages. Yufei Huo was responsible for reinforcement learning training. Yunsheng Jiang, Hao Li, and Yinghang Song contributed to the data infrastructure. Jianzhu Guo led the overall project direction and supervision. The remaining authors contributed through technical discussions and feedback.

Acknowledgments. We thank Zhuowei Chen, Gen Li, and other colleagues for their valuable discussions, suggestions, and support on Lance.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[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] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

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

[7] Shengqu Cai, Eric Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. arXiv preprint arXiv:2411.18616, 2024.

[8] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951, 2025.

[9] 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.

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7310–7320, 2024.

[11] 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.

[12] 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.

[13] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. arXiv preprint arXiv:2411.18211, 2024.

[14] 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.

[15] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences, 67(12):220101, 2024.

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

[17] Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. arXiv preprint arXiv:2509.06818, 2025.

[18] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. arXiv preprint arXiv:2507.05595, 2025.

[19] 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.

[20] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems, 36:49250–49267, 2023.

[21] Wenxun Dai, Zhiyuan Zhao, Yule Zhong, Yiji Cheng, Jianwei Zhang, Linqing Wang, Shiyi Zhang, Yunlong Lin, Runze He, Fellix Song, et al. Chatumm: Robust context tracking for conversational interleaved generation. arXiv preprint arXiv:2602.06442, 2026.

[22] 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.

[23] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. NIPS, 34:19822–19835, 2021.

[24] 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.

[25] 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 ICML, 2024.

[26] 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.

[27] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. arXiv preprint arXiv:2408.14023, 2024.

[28] Kailai Feng, Yuxiang Wei, Bo Chen, Yang Pan, Hu Ye, Songwei Liu, Chenqian Yan, and Yuan Gao. Dreamlite: A lightweight on-device unified model for image generation and editing. arXiv preprint arXiv:2603.28713, 2026.

[29] Fengyi Fu, Lei Zhang, Mengqi Huang, and Zhendong Mao. Feededit: Text-based image editing with dynamic feedback regulation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 2661–2670, 2025.

[30] Fengyi Fu, Mengqi Huang, Lei Zhang, and Zhendong Mao. Layeredit: Disentangled multi-object editing via conflict-aware multi-layer learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pages 4003–4011, 2026.

[31] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. Visual Intelligence, 2(1):1–17, 2024.

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

[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] Google DeepMind. Gemini 3 Pro Image Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf, November 2025. Model card published: November 2025.

[35] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023.

[36] Xiaochuang Han, Youssef Emad, Melissa Hall, John Nguyen, Karthik Padthe, Liam Robbins, Amir Bar, Delong Chen, Michal Drozdzal, Maha Elbayad, et al. Tv2tv: A unified framework for interleaved language and video generation. arXiv preprint arXiv:2512.05103, 2025.

[37] 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.

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

[39] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. NIPS, 33:6840–6851, 2020.

[40] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868, 2022.

[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] Mengqi Huang, Zhendong Mao, Penghui Wang, Quan Wang, and Yongdong Zhang. Dse-gan: Dynamic semantic evolution generative adversarial network for text-to-image generation. In Proceedings of the 30th ACM International Conference on Multimedia, pages 4345–4354, 2022.

[43] Mengqi Huang, Zhendong Mao, Zhuowei Chen, and Yongdong Zhang. Towards accurate image coding: Improved autoregressive image generation with dynamic vector quantization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22596–22605, 2023.

[44] Mengqi Huang, Zhendong Mao, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom: Narrowing real text word for real-time open-domain text-to-image customization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7476–7485, 2024.

[45] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. Advances in Neural Information Processing Systems, 38:167283–167308, 2026.

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

[47] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 17191–17202, 2025.

[48] Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Nanxuan Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, et al. Editverse: Unifying image and video editing and generation with in-context learning. arXiv preprint arXiv:2509.20360, 2025.

[49] Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. arXiv preprint arXiv:2503.19907, 2025.

[50] Kling AI. Kling ai. https://klingai.kuaishou.com/, 2024. Accessed: 2024-06-06.

[51] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125, 2023.

[52] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. arXiv preprint arXiv:2403.14468, 2024.

[53] Black Forest Labs. Flux: Official inference repository for flux.1 models, 2024. URL https://github.com/black-forest-labs/flux. Accessed: 2025-02-07.

[54] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742, 2025.

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

[56] 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.

[57] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems, 36:30146–30166, 2023.

[58] 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.

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

[60] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. Science China Information Sciences, 68(10):200102, 2025.

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

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

[63] 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.

[64] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In Proceedings of the 2024 conference on empirical methods in natural language processing, pages 5971–5984, 2024.

[65] 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.

[66] Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14994–15004, 2025.

[67] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. arXiv preprint arXiv:2412.06264, 2024.

[68] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

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

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

[71] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26296–26306, 2024.

[72] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

[73] Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C Pérez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, et al. Mardini: Masked autoregressive diffusion for video generation at scale. arXiv preprint arXiv:2410.20280, 2024.

[74] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. Advances in neural information processing systems, 38:40783–40818, 2026.

[75] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In European Conference on Computer Vision, pages 1–18. Springer, 2024.

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

[77] 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.

[78] 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, and Yuren Cong. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. arXiv preprint arXiv:2604.24763, 2026.

[79] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. arXiv preprint arXiv:2502.10248, 2025.

[80] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7739–7751, 2025.

[81] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), 2024.

[82] Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom++: Representing images as real-word for real-time customization. arXiv preprint arXiv:2408.09744, 2024.

[83] Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realcustom++: Representing images as real textual word for real-time customization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[84] Zhendong Mao, Mengqi Huang, Yijing Lin, Quan Wang, Lei Zhang, and Yongdong Zhang. Toward accurate image generation via dynamic generative image transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

[85] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. arXiv preprint arXiv:2504.16915, 2025.

[86] OpenAI. Gpt-4v(ision) system card. https://openai.com/index/gpt-4v-system-card/, September 2023. Accessed: 2026-05-15.

[87] OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025. Accessed: 2026-04-10.

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

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

[90] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in 200k. arXiv preprint arXiv:2503.09642, 2025.

[91] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In ICLR, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

[92] 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.

[93] 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.

[94] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In ICML, pages 8821–8831. Pmlr, 2021.

[95] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684–10695, 2022.

[96] Runway. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha, June 2024. Accessed: 2024-06-17.

[97] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. arXiv preprint arXiv:2604.14148, 2026.

[98] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427, 2025.

[99] Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation. arXiv preprint arXiv:2507.06119, 2025.

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

[101] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

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

[103] Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qi Wei, Jinhui Yin, Xue Yang, Zhihang Zhong, Qi Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, Gen Luo, Hongsheng Li, Yu Qiao, Kai Chen, and Hongjie Zhang. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. 2026. URL https://arxiv.org/abs/2603.09877.

[104] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems, 37:84839–84865, 2024.

[105] Rui Tian, Mingfei Gao, Haiming Gang, Jiasen Lu, Zhe Gan, Yinfei Yang, Zuxuan Wu, and Afshin Dehghan. Unigen-1.5: Enhancing image generation and editing through reward unification in reinforcement learning. arXiv preprint arXiv:2511.14760, 2025.

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

[107] 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.

[108] Yijing Tu, Shaojin Wu, Mengqi Huang, Wenchuan Wang, Yuxin Wang, Chunxiao Liu, and Zhendong Mao. Stream-t1: Test-time scaling for streaming video generation. arXiv preprint arXiv:2605.04461, 2026.

[109] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

[110] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 21612–21622, 2025.

[111] 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.

[112] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571, 2023.

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

[114] 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.

[115] 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.

[116] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision, 133(5):3059–3078, 2025.

[117] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In European conference on computer vision, pages 396–416. Springer, 2024.

[118] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for instruction-followed understanding and safety-aware generation. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 3907–3916, 2024.

[119] 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.

[120] Bin Wu, Mengqi Huang, Shaojin Wu, Weinan Jia, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Stream-r1: Reliability-perplexity aware reward distillation for streaming video generation. arXiv preprint arXiv:2605.03849, 2026.

[121] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870, 2025.

[122] 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.

[123] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 12966–12977, 2025.

[124] 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.

[125] Shaojin Wu, Fei Ding, Mengqi Huang, Wei Liu, and Qian He. Vmix: Improving text-to-image diffusion model with cross-attention mixing control. arXiv preprint arXiv:2412.20800, 2024.

[126] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. arXiv preprint arXiv:2508.18966, 2025.

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

[128] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In Forty-first International Conference on Machine Learning, 2024.

[129] Xiaojun Wu, Dixiang Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. Taiyi-diffusion-xl: Advancing bilingual text-to-image generation with large vision-language model support, 2024.

[130] 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.

[131] Teng Xiao, Zuchao Li, and Lefei Zhang. Omnibridge: Unified multimodal understanding, generation, and retrieval via latent space alignment. arXiv preprint arXiv:2509.19018, 2025.

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

[133] 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.

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

[135] 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.

[136] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning, 2024.

[137] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818, 2025.

[138] Chenyu Yang, Xizhou Zhu, Jinguo Zhu, Weijie Su, Junjie Wang, Xuan Dong, Wenhai Wang, Bin Li, Jie Zhou, Yu Qiao, et al. Vision model pre-training on interleaved image-text data via latent compression learning. Advances in Neural Information Processing Systems, 37:23912–23938, 2024.

[139] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670, 2025.

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

[141] Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216, 2025.

[142] 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.

[143] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision, 133(4):1879–1893, 2025.

[144] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

[145] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199, 2023.

[146] Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. arXiv preprint arXiv:2505.02567, 2025.

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

[148] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. arXiv preprint arXiv:2509.16117, 2025.

[149] 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.