UniDDT: Unifying Multimodal Understanding and Generation with Decoupled Diffusion Transformer

Shuai Wang[1] Liang Li[2] Yang Chen[1]
Ruopeng Gao[1] Yao Teng[3] Limin Wang [1,]

[1]Nanjing University [2]ByteDance Seed [3]University of Hong Kong

https://github.com/MCG-NJU/UniDDT

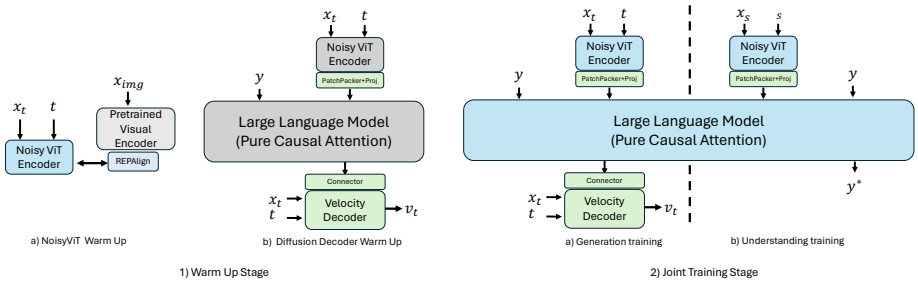


Fig. 1: The two-stages Training Recipe of UniDDT: Warmup training stage and Joint training stage. In warmup training, we warm up the Noisy ViT encoder and diffusion decoder separately to avoid model collapse caused by direct joint training. In joint training, we unfreeze all modules and optimize them through a duality of generation and understanding.

Abstract. Unified Multimodal Models (UMMs) have emerged as a critical direction for general-purpose multimodal intelligence, integrating understanding and generation into a single framework. However, existing UMMs face prominent challenges: (1) the inherent learning conflicts between visual understanding and generation tasks, leading to suboptimal modeling in both tasks; (2) different understanding and generation visual spaces impeding scalability; (3) over-reliance on task-specific data that neglects the duality of text-image understanding and generation. To address these challenges, we propose UniDDT, which leverages a Noisy ViT encoder along with a LLM to unify semantic encoding for visual generation and understanding tasks, while employing a separate diffusion decoder to decouple diffusion decoding from text decoding. With this Noisy ViT encoder, UniDDT is able to leverage the latent space as a unified visual representation, enabling seamless compatibility between understanding and generation tasks. Thus, the scalability within the generation tasks and the semantic expressiveness within understanding tasks can be balanced. Also, we construct dual data structures from the same image-text pairs, fostering interdependence between the generation and

: Corresponding author (lmwang@nju.edu.cn).

understanding data to exploit their inherent duality. Extensive experiments demonstrate that UniDDT achieves effective unification of multimodal understanding and generation with enhanced semantic consistency and scalability. For visual generation tasks, our UniDDT achieves 0.86 GenEval score and 86.9 DPG overall score. For multimodal understanding tasks, our UniDDT achieves 1699.5 score on MME benchmark and 76.5 overall score on SEEDbench.

Fig. 2: The curated samples (Max Resolution 1024×1024**) from UniDDT.** We adopt Adams-2nd solver with 25 steps and CFG value of 4. We place the respective prompts and more visual samples in Appendix.

1 Introduction

Unified Multimodal Models (UMMs) [17, 34, 38, 39, 48, 62, 91] integrate both understanding and generation into a single framework. Numerous initiatives [17, 34, 38, 79] have thrived to close the gap with proprietary unified multimodal systems. Following recent iterative refinement, mainstream text-image UMMs now largely adopt a hybrid AR-diffusion approach (autoregressive for text generation and diffusion for visual generation). However, due to the inherent substantial differences between understanding and visual generation, the implementation of AR-diffusion hybrid UMMs is still neither definitive nor straightforward. We will elaborate on this from three key perspectives: modeling, visual space, and training data.

From Modeling Perspective: UMMs were initially envisioned to achieve mutual promotion between understanding and generation. Early attempts, particularly hybrid AR-diffusion models, used adapters [39, 48] to assemble task-specific tailored models. However, these assembly-based approaches represent a relatively

superficial integration, failing to fully exploit the potential synergy between understanding and generation. An alternative approach natively integrates both objectives within a single framework, deemed as Native-UMMs. Mainstream Native-UMMs typically adopt parallel branches for understanding and generation, yet this often results in a significant performance trade-off, still failing to demonstrate the hypothesized mutual promotion. To mitigate this conflict, these UMMs [17, 37, 38] often decouple the parameters specific to each task.

From the Visual Space Standpoint: A consensus has not yet been reached on the optimal unified visual space. Understanding models thrive on high-dimensional semantic representations, while generative models struggle with training in such spaces. Consequently, most UMMs [17, 38] adopt distinct visual spaces for different tasks (e.g., a semantic space for understanding and a VAE space for visual generation). This inherent decoupling results in fragmented visual spaces, complicates the overall workflow and impedes large-scale scaling. In response, Unified Visual Space UMMs [34, 79] employ a single, shared space. Specifically, some works [90] opt for the semantic-rich representations of visual foundation models [47, 67, 88], while others [79] choose the detail-rich VAE latent space. Moreover, raw pixel space [18, 69] seems more scalable, but not validated.

From the Training Data Viewpoint: although Mogao [38] hypothesizes that interleaved multi-modal training data is the key to true unification, most existing UMMs do not follow this approach. Instead, they effectively stitch understanding and generation components together by training each on task-specific data.

To address the aforementioned problems, we propose UniDDT, a native UMM that features a unified visual space and a decoupled but unified understanding-generation design. An architectural comparison between UniDDT and other UMMs is provided in Fig. 3. Although most UMMs are broadly built upon three core components—diffusion transformers, VLM-based semantic encoders, and hybrid autoregressive–diffusion frameworks—the specific architectural designs differ substantially across models. UniDDT leverages a Noisy ViT encoder along with a LLM to unify semantic encoding for visual generation and understanding tasks, while employing a separate diffusion decoder to decouple diffusion decoding from text decoding. We argue that the key contribution of UniDDT lies in this unified semantic encoding paradigm combined with the explicit separation of semantic modeling and high-frequency visual decoding.

As shown in Fig. 1, we treat visual understanding as a preceding task for visual generation. This allows the Noisy ViT encoder and LLM backbone to unify two key processes in modeling standpoint: the semantic extraction for standard visual understanding and the semantic encoding of the Noisy inputs for diffusion-based generation. The separate diffusion decoder is then dedicated to visual generation, conditioned on the semantics encoded by the ViT and LLM backbone. Regarding the visual space, the Noisy ViT encoder enables the unification of visual spaces, thus we compared pixel and latent spaces choices. Although pixel space holds a slight advantage for understanding, it suffers from a significant deficit in generative performance and does not demonstrate superior scaling properties. Therefore, we adopt the latent space as our principal visual space.

From the training data viewpoint, we leverage the understanding-generation duality to enhance our UniDDT under limited data scale.

Our contributions are summarized as follows:

- We propose UniDDT, a native UMM with a unified visual space and a decoupled but unified understanding-generation design.
- Our VLM-UniDDT achieves 0.87 overall score on GenEval benchmark and 86.9 score on DPG benchmark, meanwhile, it achieves 1699.5 perception score on MME Benchmark and 76.5 overall score on SEEDbench.

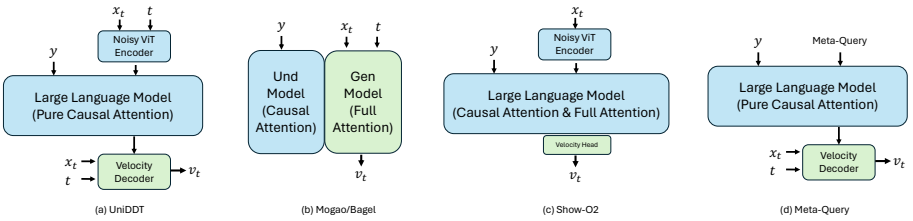


Fig. 3: Architecture comparison with UniDDT and other unified multimodal models. UniDDT consists of a noisy vit encoder, an llm backbone, and a diffusion decoder. The noisy vit encoder and LLM backbone unify the semantic perception of understanding and generation. The diffusion decoder is dedicated to visual generation.

2 Related Works

Visual Language Models. Modern Visual Language Models (VLMs) [2, 3, 13, 68, 73, 93] are built on LLMs and trained under the classic next-token prediction paradigm. These VLMs leverage pre-trained visual encoders [67, 88] to align raw pixels with the language embedding space. Early attempts explored raw-pixel inputs [18], and causal discrete visual tokens [35, 92], but yielded inferior performance. Other [63, 65] attempts combined different visual encoders for fine-grained perception.

Visual Generative Models. High-performance generative models [23, 26, 58] typically rely on latent diffusion models. Modern latent diffusion models comprise a variational autoencoder (VAE) [8, 9, 56, 80, 82, 83] and a diffusion model [4, 43, 51, 70, 71], trained on a latent space shaped by the VAE. Under the classic latent diffusion setup, researchers have leveraged pre-trained visual foundation models to boost performance. Some works [86] adopt visual foundation models to align the intermediate features of the diffusion model, while others [80, 83] align the VAE's latent space. Additionally, DDT [71] enhances generative capability through decoupling diffusion transformer into a tailored architecture. Instead, More ambitiously, RAE [90] dispenses the traditional VAE and some [11, 69] strikes back to pixel space. Current discrete visual generation [1, 20, 46, 61, 85] still performs inferior.

Unified Multimodal Models. Inspired by the success of large language models, discrete token-based unified models [12, 15, 24, 29, 42, 60, 62, 72, 76, 78] convert pixels into discrete visual tokens and are then trained under a unified next-token-prediction paradigm. To mitigate the loss in generative performance in pure discrete auto-regressive approach, AR-diffusion hybrid approaches [17, 37, 38, 44, 74, 79, 91] leverage discrete auto-regressive modeling for text generation and diffusion modeling for image generation. Beyond native unified frameworks, an alternative research direction [6, 48] involves integrating specialized large multimodal models with diffusion-based generative models by tuning adapters. Concurrently, Representation-Forcing [74] proposes to bridge the representation gap across modalities using semantic autoregressive tokens. RepFusion [49] proposes a unified architecture similar to ours, but further equips it with a powerful RAE [59, 66, 90]. However, current open-source unified multimodal models still exhibit a significant gap compared to proprietary systems.

3 Method

3.1 Revisit DDT Architecture

As shown in Fig. 3, DDT [71] consists of a heavy condition encoder and a light velocity decoder. The heavy condition encoder takes three inputs, a prompt condition $\boldsymbol{y}$, noisy inputs $\boldsymbol{x}_t$, and timestep t, to extract the self-condition feature $\boldsymbol{z}_t$ through stacked diffusion transformer blocks.

$$\boldsymbol{z}_t = \textbf{Encoder}\ (\boldsymbol{x}_t, t, y). \tag{1}$$

DDT adopts the representation alignment technique from REPA [86] and aligns the intermediate feature $\mathbf{h}_i$ from the i-th layer in the condition encoder with the DINOv2 representation r_*. Consistent to REPA [86], the h_ϕ is the learnable projection MLP:

$$\mathcal{L}_{enc} = 1 - \cos(r_*, h_\phi(\mathbf{h_i})). \tag{2}$$

The velocity decoder mirrors the encoder model architecture; it takes the noisy latent $\boldsymbol{x}_t$, timestep t, and self-conditioning $\boldsymbol{z}_t$ as inputs to estimate the velocity $\boldsymbol{v}_t$. The external-condition timestep t and self-condition feature $\boldsymbol{z}_t$ are used as conditions for the decoder blocks:

$$\boldsymbol{v}_t = \textbf{Decoder}\ (\boldsymbol{x}_t, t, \boldsymbol{z}_t). \tag{3}$$

The velocity decoder is trained with the flow matching loss as shown in Eq. (4):

$$\mathcal{L}_{dec} = \mathbb{E}[\int_0^1 ||(\boldsymbol{x}_{data} - \epsilon) - \boldsymbol{v}_t(\boldsymbol{x}_t, t, \boldsymbol{z}_t|\theta)||^2 \mathrm{d}t]. \tag{4}$$

Finally, DDT jointly trains the condition encoder and the velocity decoder with Eq. (2) and Eq. (4). So far, the DDT-like arch has been adopted in RAE [90] and PixNerd [69].

3.2 UniDDT Architecture

UniDDT adheres to the core philosophy of DDT and is tailored for the **unified understanding and generation** of text and images. Specifically, UniDDT comprises three key components: a Noisy ViT encoder, a large language model (LLM) backbone, and a diffusion decoder. Oracle DDT comprises a diffusion decoder and a condition encoder. In UniDDT, the NoisyViT encoder together with the LLM backbone forms the condition encoder module.

As shown in Fig. 1, the noisy vit encoder takes the noisy input $\boldsymbol{x}_t$ and timestep t as inputs to extract high-level semantics $\boldsymbol{s}_t$. If we take the pixel space as the unified visual space, $\boldsymbol{x}_t$ is the noisy image. If we take the latent space as the unified visual space, $\boldsymbol{x}_t$ refers to the noisy latent. The details about unified visual space can be found in Sec. 3.3. For multimodal understanding, the LLM backbone first performs causal encoding on $\boldsymbol{s}_t$, then autoregressively decodes $\boldsymbol{y}$. For visual generation, the LLM backbone causally processes the prompt condition $\boldsymbol{y}$ and visual semantics $\boldsymbol{s}_t$, injecting the semantics from $\boldsymbol{y}$ into $\hat{\mathbf{s}}_t$. The diffusion decoder takes the refined visual semantics $\hat{\mathbf{s}}_t$ (derived from $\boldsymbol{s}_t$) as the condition, then estimates the velocity $\boldsymbol{v}_t$ from the noisy input $\boldsymbol{x}_t$. Below, we elaborate on each component in detail:

Noisy ViT Encoder. Our noisy vit encoder mirrors the architecture design as the condition encoder of DDT [71]. It is built with interleaved Attention and FFN blocks. The encoder processes two inputs, the noisy latent $\boldsymbol{x}_t$ and timestep t to extract the semantic feature $\boldsymbol{z}_t$ through a series of stacked Attention and FFN blocks:

$$\boldsymbol{z}_t = \textbf{NoisyViT}\ (\boldsymbol{x}_t, t). \tag{5}$$

Similar to DiT [51] and SiT [43], we inject the timestep condition through AdaLN-zero [51].

LLM Backbone. Following the common practice of Qwen-VL [2], we construct distinct chat templates for understanding and visual generation tasks, replacing the image placeholder token with the corresponding image semantic features $\boldsymbol{z}_t$ extracted by the Noisy ViT. For multimodal understanding, the LLM backbone first causally encodes the visual semantics $\boldsymbol{z}_t$ and understanding prefix tokens (denoted as $\boldsymbol{y}$), then autoregressively decodes new text tokens $\boldsymbol{y}^*$:

$$\boldsymbol{y}^* = \mathrm{LLM}\left(\boldsymbol{z}_t, \boldsymbol{y}\right). \tag{6}$$

For the visual generation task, the LLM backbone causally encodes the prompt condition $\boldsymbol{y}$ and visual semantics $\boldsymbol{z}_t$ to perform semantic injection, yielding refined $\hat{\boldsymbol{z}}_t$. These refined visual features $\hat{\boldsymbol{z}}_t$ are then fed into the diffusion decoder:

$$\hat{\boldsymbol{z}}_t = \mathrm{LLM}\left(\boldsymbol{y}, \boldsymbol{z}_t\right). \tag{7}$$

Diffusion Decoder. It adopts the same architectural framework as the Noisy ViT encoder, comprising stacked interleaved Attention and FFN blocks—similar to DiT/SiT. It takes the noisy latent $\boldsymbol{x}_t$, timestep t as inputs, and $\hat{\boldsymbol{z}}_t$ as a condition to estimate the velocity $\boldsymbol{v}_t$. Through experiments, we found that training the diffusion decoder using only $\hat{\boldsymbol{z}}_t$ (without text tokens) is feasible, even when the LLM backbone and Noisy ViT encoder are frozen. Unlike DDT [71], we use attention (instead of AdaLN-zero) to inject the $\hat{\boldsymbol{z}}_t$ condition into the diffusion decoder features. As shown in Eq. (8), we elaborate on the diffusion decoder's block structure. For readability, we retain the notation $\mathbf{x}_t$ for the intermediate feature:

$$\mathbf{x}_t = \mathbf{x}_t + \textsc{AdaLN}\left(\boldsymbol{t}, \textsc{Attention}\left(\mathbf{x}_t, \hat{\boldsymbol{z}}_t\right)\right), \tag{8}$$

$$\mathbf{x}_t = \mathbf{x}_t + \textsc{AdaLN}\left(\boldsymbol{t}, \mathrm{FFN}\left(\mathbf{x}_t\right)\right). \tag{9}$$

To improve the training stability, we add several full attention transformer blocks as the refiner [21, 69] to refine the provided last hidden states.

3.3 Unified Visual Space

Previous understanding model, typically, VLMs [2, 3] takes the raw pixels as the principal visual space. While generative models rely on a largely compressed latent space [8, 56, 80, 83] to eliminate redundancy and ease generative model learning. Thus, there exists a learning space tradeoff for a unified model. We found that the understanding performance of the latent space is marginally inferior to that of the pixel space, the performance degradation is minimal, so they can be regarded as comparable in terms of understanding. When it comes to generation performance, though, the latent space is significantly better than the pixel space, and no better scaling advantage has been identified for the pixel space compared to the latent space in our experiments. This motivates us to take the latent space as the principal visual space of UniDDT.

3.4 UniDDT Training

Warmup Training. Starting joint training from random initialization can easily cause language model collapse; thus, we employ a separate warmup stage. We use a pre-trained vision-language model (VFM), e.g., SigLIP [67, 88] or Qwen3-ViT [81], as the teacher model to distill representations to the Noisy ViT encoder. Once the Noisy ViT encoder converges, we freeze its parameters along with those of the LLM backbone, then warm up the diffusion decoder.

Specifically, for the Noisy ViT encoder, we initialize most of its parameters from the teacher model, except for the timestep AdaLN-Zero modules. Note that different from show-o2 [79], our Noisy ViT encoder also takes t as a extra condition for semantic extraction. We then distill representations from the teacher model to the Noisy ViT encoder as defined in Eq. (2).For the diffusion decoder warmup (as shown in Fig. 1), we add a projection layer to align the dimensions of the Noisy ViT and LLM backbone if needed. We freeze the Noisy

ViT encoder and LLM backbone, and jointly train this projection layer and the diffusion decoder using the flow-matching loss specified in Eq. (4).

Joint Training. Previous unified models used distinct image-text pairs for understanding and generation tasks. We argue that understanding and generation can be framed as a dual task and leverage this duality to boost our UniDDT under limited data scale. Specifically, given a text-image pair $(\mathbf{y}, \mathbf{x})$, we construct the data formats as follows (the actual template is provided in the Appendix):

<user>generate.$\boldsymbol{y}$<user><assistant>$\boldsymbol{x}$<assistant>

The understanding-oriented data is constructed as:

<user>describe.$\boldsymbol{x}$<user><assistant>$\boldsymbol{y}$<assistant>

During this stage, we unfreeze all modules to initiate training: we randomly sample between the understanding and generation formats, apply only the cross-entropy loss to the text $\boldsymbol{y}$ in the understanding task, and use the diffusion loss for $\boldsymbol{x}$ in the generation task. Through experiments, we find this joint training stage benefits visual generation a lot. The joint loss is defined as:

$$\mathcal{L}_{joint} = \mathbb{E}_{\text{gen}}\mathcal{L}_{diff}(\boldsymbol{x}|\boldsymbol{y}) + \lambda_{\text{und}}\mathbb{E}_{\text{und}}\mathcal{L}_{\text{ce}}(\boldsymbol{y}|\boldsymbol{x}). \tag{10}$$

Post Training. After the joint training stage, UniDDT can not only generate novel images but also understand the intermediate states of the generation process. This inspires us to further enhance generation quality by leveraging this unique property. Specifically, we freeze the Noisy ViT encoder and LLM backbone, and only train the diffusion decoder during the post-training stage. Given a noisy input $\boldsymbol{x}_t$, its corresponding timestep t, and the prompt $\boldsymbol{y}$, UniDDT takes these three inputs and yields the estimated velocity $\boldsymbol{v}_t$. A noisy point at time s can be estimated as: $\boldsymbol{x}_s = \boldsymbol{x}_t + \boldsymbol{v}_t(s-t)$.

$$\boldsymbol{x}_s = \boldsymbol{x}_t + \boldsymbol{v}_t(\boldsymbol{x}_t, t|\theta)(s-t). \tag{11}$$

We then feed the intermediate states $\{\boldsymbol{x}_s, s\}$ into UniDDT's understanding branch to estimate the likelihood $\log p(\boldsymbol{y}|\boldsymbol{x}_s, s)$, and maximize this likelihood to improve generation quality and semantic consistency:

$$\mathcal{L}_{post} = \mathbb{E}_{\boldsymbol{x},t,s,\boldsymbol{y}}\mathcal{L}_{\text{ce}}(\boldsymbol{y}|\boldsymbol{x}_s, s). \tag{12}$$

4 Experiments

Dataset. We collect a mixed dataset with approximately 70M images from publicly available datasets [5, 16, 30, 45, 52, 57]. We recaption every image with Qwen2.5-VL-7B [3] to yield captions with various lengths. The data sources and caption details can be found in the appendix. We place the detailed chat templates for generation and understanding in the Appendix.

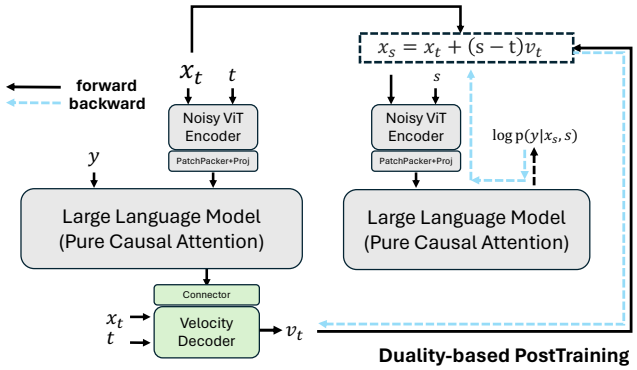


Fig. 4: The duality-based post-training of UniDDT. We freeze the understanding related components and only unfreeze the diffusion decoder; then we feed the intermediate results of visual generation to understanding branch to maximize the likelihood.

Model Details. We provide detailed model specifications in Tab. 1. We instantiate two UniDDT variants, differing in their backbone architectures: one with an LLM backbone (denoted as NativeUniDDT) and the other with a VLM backbone (denoted as VLMUniDDT). As shown in Tab. 1, NativeUniDDT-B is configured with a 12-layer Noisy ViT encoder, Qwen3-0.6B[1] as its LLM backbone, and a 20-layer diffusion decoder with 1024 dimensions(4-layer refiner). NativeUniDDT-L features a 24-layer, 1024-dimension Noisy ViT encoder, a 20-layer, 1536-dimension diffusion decoder, and adopts Qwen3-1.7B[2] as its LLM backbone. For NativeUniDDT-XL, we scale the diffusion decoder dimension of NativeUniDDT-L to 2560. VLM-UniDDT adopts Qwen3-VL-4B[3] as the LLM backbone, while other components are configured consistently with NativeUniDDT-L. We adopt the latent space of Flux-VAE[4] as the unified visual space of UniDDT.

Table 1: The detailed model configurations of UniDDT. UniDDT has two distinct variant, differing in their LLM backbones. NativeUniDDT adopts Qwen3 as its LLM backbone while VLM-UniDDT uses Qwen3-VL as its llm backbone.

Config	LLM	NoisyViT		Diffusion Decoder		
		#Layers	#Dim	#Layers	#Heads	#Dim
NativeUniDDT-B	Qwen3-0.6B	12	768	16+4	16	1024
NativeUniDDT-L	Qwen3-1.7B	24	1024	16+4	24	1536
NativeUniDDT-XL	Qwen3-1.7B	24	1024	16+4	20	2560
VLM-UniDDT	Qwen3VL-4B	24	1024	16+4	24	1536

[1] https://huggingface.co/Qwen/Qwen3-0.6B
[2] https://huggingface.co/Qwen/Qwen3-1.7B
[3] https://huggingface.co/Qwen/Qwen3-VL-4B
[4] https://huggingface.co/diffusers/FLUX.1-vae

Training Details. To avoid the mismatch of training text-image pairs caused by center crops, we adopt native aspect ratio training [75]. We adopt FSDP [50,89] to shard model parameters, eliminating memory redundancy. For the Native-UniDDT, we choose SigLIP2 [67] as the teacher model of the noisy vit encoder. SigLIP2-B for NativeUniDDT-B, SigLIP2-so-400M for Native-UniDDT-L and Native-UniDDT-XL, respectively. For VLM-UniDDT, we adopt the original visual encoder(Qwen-NaViT [81]) as the teacher model for the noisy vit encoder. In the warmup stage, we train the noisy vit encoder for 40K steps with a constant learning rate of 2e-4 and ema rate of 0.9999. After obtained the well-initialized noisy ViT encoder, we add a proj layer to align the noisy vit dimension to the LLM backbone, and jointly train the diffusion decoder and the aforementioned proj layer for 100K steps at maximal sequence length of 16384. For the joint training stage, we unfreeze all modules and optimize with a maximal sequence of 8192 (120K steps for Native-UniDDT, 10K steps for VLM-UniDDT). In order to further enhance the performance, we follow the common practice [6] to fine-tuning our UniDDT on OpenAI-4o datasets [6,14,84] for 8K steps. Our default training hardware consists of 16× A100.

4.1 Visual Space

We adopt the latent space of Flux-VAE and the raw pixel space as the candidate visual spaces. The latent space of Flux-VAE has 16 channels with a down-sample factor of 8. Our findings reveal that the pixel space exhibits a slight advantage over the latent space in understanding, though this is accompanied by poorer scaling properties in visual generation during the pretraining stage.

Understanding Perspective. We collected understanding metrics under the VLM-UniDDT setting, which enabled us to readily obtain cosine similarity and multimodal understanding metrics. The cosine similarity reflects, to some extent, the potential for multimodal understanding. As shown in Fig. 6a, we fed clean images into the teacher model and Noisy images with varying noise timesteps into the Noisy ViT encoder, then calculated the similarity between their features. The Noisy ViT in pixel space exhibited slightly better and more stable similarity across different noise levels, though the performance gap was negligible. The cosine similarity for Native-UniDDT followed a similar trend. For multimodal understanding metrics, we replaced the original visual encoder of Qwen3-VL-4B with our Noisy ViT encoder to collect multimodal performance data. As shown in Fig. 6d, the Noisy ViT encoder in pixel space still performed slightly better.

Generation Perspective. We collected generation metrics under the Native-UniDDT setting. Across all training stages and spaces, visual generation performance exhibited a clear scaling property. In the warmup and joint training stages, show in Fig. 6k and Fig. 6j, the pixel space did not demonstrate better scaling properties than the latent space. In the post-training stage, shown in Fig. 6l, the pixel space appeared to perform better, yet the performance gap still remained significant.

4.2 Multimodal Understanding

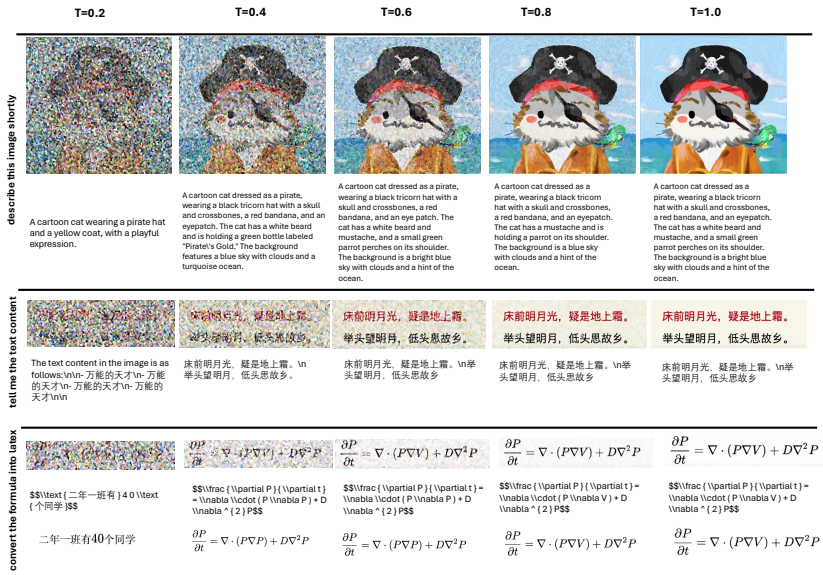


Fig. 5: The understanding power of UniDDT on Noisy inputs. VLM-UniDDT understands inputs very well under acceptable noise levels. The left vertical axis represents the user prompt, while the bottom axis shows the generated results.

Note that the timesteps of the understanding training in joint training stage is randomly sampled from $[0, 1]$, thus UniDDT is capable of understanding noisy image under reasonable noise level. We provide the understanding power of UniDDT on Noisy inputs in Fig. 5. Our model understands input images very well under acceptable noise levels.

Since our data-source only consists of naive image-text pairs, our Native-UniDDT is only capable of caption images and not follows the instructions, thus we decide not to include the understanding metrics of Native-UniDDT in Tab. 2. We collect the understanding performance on MME [22], SEED-bench [33], MMMU [87], MMStar [10], AI2D [31] benchmarks. We directly fix the timestep $t = 1.0$ for multimodal understanding evaluation. As shown in Tab. 2, our VLM-UniDDT initialized from Qwen3-VL-4B, achieves superior understanding performance across different benchmarks.

4.3 Visual Generation

Shown in Tab. 3, we provide the final performance after finetuned on 4o-like datasets [6,14,84]. Our Native-UniDDT-L achieves 0.88 overall Score on Geneval and 86.6 on DPG benchmark. Our Native-UniDDT-XL achieves 0.89 overall Score on GenEval [25] and 87.1 on DPG [27] benchmark. Our VLM-UniDDT

Table 2: Comparison with others on multimodal understanding benchmarks.

Models	# Params.	MME(p) ↑	GQA↑	SEED↑	MMB(en)↑	MMMU(val) ↑	MMStar ↑	AI2D ↑
Understanding. Only								
LLaVA-v1.5 [41]	7B	1510.7	62.0	58.6	64.3	-	-	-
Qwen-VL [2]	7B	1487.6	57.5	58.2	60.6	-	-	57.7
LLaVA-OV [32]	7B	1580.0	-	-	80.8	48.8	57.5	81.4
Unified Models								
MetaMorph [64]	8B	-	-	71.8	75.2	-	-	-
TokenFlow-XL* [53]	14B	1551.1	62.5	72.6	76.8	43.2	-	75.9
ILLUME [28]	7B	1445.3	-	72.9	75.1	38.2	-	71.4
BAGEL [17]	14B	1687.0	-	-	85.0	55.3	-	-
Show-o [78]	1.3B	1097.2	58.0	51.5	-	27.4	-	-
JanusFlow [44]	1.5B	1333.1	60.3	70.5	74.9	29.3	-	-
Janus-Pro [76]	1.5B	1444.0	59.3	68.3	75.5	36.3	-	-
Emu3 [72]	8B	-	60.3	68.2	58.5	31.6	-	70.0
VILA-U [40]	7B	1401.8	60.8	59.0	-	-	-	-
Liquid [77]	8B	1448.0	61.1	-	-	-	-	-
Janus-Pro [76]	7B	1567.1	62.0	72.1	79.2	41.0	-	-
Mogao [38]	7B	1592.0	60.9	74.6	75.0	44.2	-	-
Show-o2 [79]	7B	1620.5	63.1	69.8	79.3	48.9	56.6	78.6
VLM-UniDDT	**4B+1B**	**1699.5**	**62.3**	**76.5**	**82.2**	**52.6**	**57.7**	**78.1**

achieves 0.87 on GenEval and 86.9 on DPG benchmark. Our UniDDT beats other generative models and unified multi-modal models with a significant margin. We also provide the visual samples in Fig. 2 and Appendix.

4.4 Ablation

Warm-up Noisy ViT Encoder. Time shift [19] and log-normal [19] play a pivotal role in training diffusion-based generative models. We also adopt these strategies during the warm-up stage of the Noisy ViT encoder. However, as shown in Fig. 6g, we surprisingly found that a commonly used large time-shift value (corresponding to more noisy timesteps) significantly impairs visual understanding performance—particularly OCR capability. This observation inspires us to use a small time-shift value, which generalizes well to more noisy steps. We also include a clean ViT variant, which is trained exclusively on clean inputs. While the clean ViT encoder has limited generalization to noisy timesteps, our Noisy ViT encoder performs consistently well across different Noisy timesteps. While perceptual performance is affected by the noise level (or timestep selection), the performance curve remains monotonic. For practical use, the cleanest timestep (corresponding to the best performance) can be directly adopted without manual tuning.

Warm-up Diffusion Decoder. Conventional diffusion transformers use the last hidden states of text tokens from the LLM as conditioning. In contrast, our diffusion decoder exclusively adopts the refined visual features from the LLM backbone. This raises doubts about whether these refined visual features can efficiently summarize the prefix text prompt. As shown in Fig. 6b and Fig. 6c, the

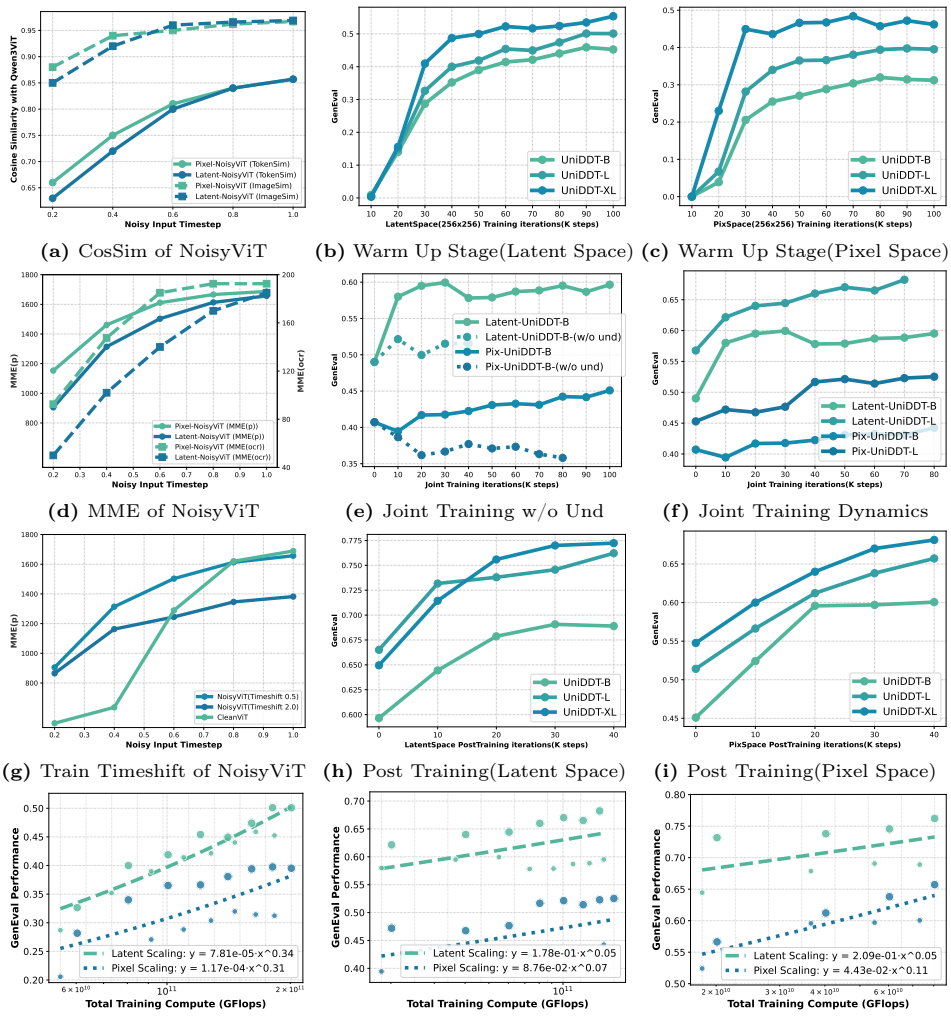
Table 3: Comparison of various methods on GenEval and DPGBench benchmarks. † indicates Generation with prompt rewriting.

METHOD	GenEval							DPGBench (↑)
	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall (↑)	
Generation. Only								
SDv1.5 [56]	0.97	0.38	0.35	0.76	0.04	0.06	0.43	63.18
SDv2.1 [56]	0.98	0.51	0.44	0.85	0.07	0.17	0.50	-
SD3-Medium [19]	0.99	0.94	0.72	0.89	0.33	0.60	0.74	84.08
SDXL [56]	0.98	0.74	0.39	0.85	0.15	0.23	0.55	74.65
PixArt-α [7]	0.98	0.50	0.44	0.80	0.08	0.07	0.48	71.11
DALL-E 2 [54]	0.94	0.66	0.49	0.77	0.10	0.19	0.52	-
DALL-E 3 [55]	0.96	0.87	0.47	0.83	0.43	0.45	0.67	83.50
Unified Models								
Chameleon [62]	-	-	-	-	-	-	0.39	-
Show-o [78]	0.98	0.80	0.66	0.84	0.31	0.50	0.68	-
Show-o2-7B [79]	1.00	0.87	0.58	0.92	0.52	0.62	0.76†	86.14
Janus [76]	0.97	0.68	0.30	0.84	0.46	0.42	0.61	79.68
JanusFlow [44]	0.97	0.59	0.45	0.83	0.53	0.42	0.63	80.09
Janus-Pro-1B [12]	0.98	0.82	0.51	0.89	0.65	0.56	0.73	82.63
Janus-Pro-7B [12]	0.99	0.89	0.59	0.90	0.79	0.66	0.80	84.19
MetaQuery-B [48]	-	-	-	-	-	-	0.74†	80.04
MetaQuery-L [48]	-	-	-	-	-	-	0.78†	81.10
MetaQuery-XL [48]	-	-	-	-	-	-	0.80†	82.05
BLIP3-o-4B [6]	-	-	-	-	-	-	0.81	79.36
BLIP3-o-8B* [6]	-	-	-	-	-	-	0.84	81.60
BAGEL-7B [17]	0.98	0.95	0.84	0.95	0.78	0.77	0.88†	-
VLM-UniDDT(512×512)	**0.99**	**0.93**	**0.71**	**0.92**	**0.85**	**0.80**	**0.87**	**86.9**

diffusion decoder learns effectively even when the Noisy ViT encoder and LLM backbone are frozen. Furthermore, as shown in Fig. 6j, performance improves steadily as the allocated training compute increases, no matter in pixel space or latent space.

Joint Training. Consistent with Sec. 3.4, we construct joint training data from the same text-image pairs by leveraging the duality between generation and understanding. To validate the effectiveness of the joint training stage for visual generation, we designed a targeted experiment: we unfreeze all modules (consistent with standard joint training) while setting the understanding loss weight to zero. As shown in Fig. 6e, this targeted experiment is denoted as *(w/o und)*. In contrast to the default duality-leveraging joint training, training exclusively on visual generation data yields inferior performance. Specifically, Latent-Native-UniDDT trained in the absence of understanding data achieves only marginal improvements, while Pixel-Native-UniDDT-B even exhibits performance degradation—which further confirms the instability of the pixel space. By contrast, under duality-aware joint training, Latent-Native-UniDDT-B delivers a significant performance leap, and Pixel-Native-UniDDT-B improves steadily. As shown in Fig. 6f and Fig. 6k, larger model sizes correspond to superior performance, with joint training exhibiting clear computational scaling behavior.

Post Training. Empowered by the duality of understanding and generation, our post-training significantly boosts generation performance. As shown in Fig. 6i and Fig. 6h, visual generation performance improves steadily. As illustrated in Fig. 6l, duality-based post-training exhibits clear scaling behavior.



(a) CosSim of NoisyViT **(b)** Warm Up Stage(Latent Space) **(c)** Warm Up Stage(Pixel Space)

(d) MME of NoisyViT **(e)** Joint Training w/o Und **(f)** Joint Training Dynamics

(g) Train Timeshift of NoisyViT **(h)** Post Training(Latent Space) **(i)** Post Training(Pixel Space)

(j) Scaling Law(Warmup Training) **(k)** Scaling Law(Joint Training) **(l)** Scaling Law(Post Training)

Fig. 6: The ablation studies table. We conduct vigorous ablation experiments on UniDDT. Specifically, we provide the training curves of warmup stage, joint training stage and the post-training stage of UniDDT.

Improvements of each stage. To better isolate the contribution of our architecture from the effects of additional fine-tuning data, we report the performance of VLM-UniDDT at each training stage prior to fine-tuning on OpenAI GPT-4o-style data.

Stage	2-Obj.	Count	Pos.	Color	Overall
Warmup	0.63	0.32	0.20	0.27	0.52
Joint	0.69	0.45	0.41	0.37	0.60
Post	0.84	0.56	0.50	0.61	0.72
+4o FT	0.93	0.71	0.85	0.80	0.87

Table: Ablation across training stages.

5 Limitation

Although we emphasize making full use of the duality of data, the text in the original image-text data mainly comes from captions generated by other models. This greatly limits the understanding ability and instruction following capability of Native-UniDDT leaving it only capable of performing image captioning tasks. Thus, we decide not to report the understanding performance of Native-UniDDT, and only provide the performance of VLM-UniDDT in Tab. 2 and Tab. 3. We believe that increasing the richness of the original data helps address this issue. Concurrent work [49] adopting the similar architecture also uses a stronger VAE, which is likewise a promising direction for further exploration and improvement. Since our experiments were conducted before the release of JiT [36], our pixel-space experiments did not consider the prediction formulation proposed by JiT, leaving room for further improvement.

6 Conclusion

Unified Multimodal Models (UMMs) are pivotal for advancing general-purpose multimodal intelligence, yet existing approaches face core challenges: conflicting objectives between understanding and generation in modeling, fragmented visual spaces that hinder scalability, and task-specific training data failing to leverage text-image duality. To address these, we propose UniDDT, a native UMM with a decoupled yet unified design—comprising a Noisy ViT encoder, an LLM backbone, and a diffusion decoder—that frames understanding as a prerequisite for generation to avoid mutual trade-offs, adopts the latent space as the unified visual representation for balanced semantic expressiveness and scalability, and constructs dual understanding-generation data from the same image-text pairs without relying on task-specific datasets. This concise framework enhances semantic consistency between understanding and generation, demonstrating that decoupling conflicting objectives, unifying visual representation, and leveraging task duality are key to advancing UMMs, and offers a new perspective for next-generation unified model design.

Acknowledgements. This work is supported by the Basic Research Program of Jiangsu (No. BK20250009), the Fundamental Research Funds for the Central Universities (No. 020214380140), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025 XDXM118), the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025) 4
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization. Text Reading, and Beyond **2**, 1 (2023) 4, 6, 7, 12
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 4, 7, 8
4. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22669–22679 (2023) 4
5. common canvas: common-canvas/commoncatalog-cc-by-nc-sa · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/common-canvas/commoncatalog-cc-by-nc-sa (2024), [Accessed 06-11-2025] 8
6. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025) 5, 10, 11, 13
7. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-\alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 13
8. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Lu, Y., Han, S.: Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733 (2024) 4, 7
9. Chen, J., Zou, D., He, W., Chen, J., Xie, E., Han, S., Cai, H.: Dc-ae 1.5: Accelerating diffusion model convergence with structured latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19628–19637 (2025) 4
10. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024) 11
11. Chen, S., Ge, C., Zhang, S., Sun, P., Luo, P.: Pixelflow: Pixel-space generative models with flow. arXiv preprint arXiv:2504.07963 (2025) 4
12. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025) 5, 13
13. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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. pp. 24185–24198 (2024) 4
14. Chen, Z., Bai, X., Shi, Y., Fu, C., Zhang, H., Wang, H., Sun, X., Zhang, Z., Wang, L., Zhang, Y., et al.: Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. arXiv preprint arXiv:2509.24900 (2025) 10, 11
15. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al.: Emu3. 5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025) 5

16. deepghs: deepghs/midjourney_captioned_23m_full · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/deepghs/midjourney_captioned_23m_full (2024), [Accessed 06-11-2025] 8
17. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 2, 3, 5, 12, 13
18. Diao, H., Cui, Y., Li, X., Wang, Y., Lu, H., Wang, X.: Unveiling encoder-free vision-language models. Advances in Neural Information Processing Systems **37**, 52545–52567 (2024) 3, 4
19. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206 (2024) 12, 13
20. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 4
21. Fan, L., Li, T., Qin, S., Li, Y., Sun, C., Rubinstein, M., Sun, D., He, K., Tian, Y.: Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. arXiv preprint arXiv:2410.13863 (2024) 7
22. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al.: Mme: a comprehensive evaluation benchmark for multimodal large language models. corr abs/2306.13394 (2023) (2023) 11
23. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025) 4
24. Gao, Z., Shou, M.Z.: D-ar: Diffusion via autoregressive models. arXiv preprint arXiv:2505.23660 (2025) 5
25. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023) 11
26. Gong, L., Hou, X., Li, F., Li, L., Lian, X., Liu, F., Liu, L., Liu, W., Lu, W., Shi, Y., et al.: Seedream 2.0: A native chinese-english bilingual image generation foundation model. arXiv preprint arXiv:2503.07703 (2025) 4
27. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024) 11
28. Huang, R., Wang, C., Yang, J., Lu, G., Yuan, Y., Han, J., Hou, L., Zhang, W., Hong, L., Zhao, H., Xu, H.: Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. arXiv preprint arXiv:2504.01934 (2025) 12
29. Jiao, Y., Qiu, H., Jie, Z., Chen, S., Chen, J., Ma, L., Jiang, Y.G.: Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3600–3610 (2025) 5
30. JourneyDB: JourneyDB/JourneyDB · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/JourneyDB/JourneyDB (2023), [Accessed 06-11-2025] 8
31. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016) 11
32. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 12

33. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023) 11
34. Li, H., Peng, X., Wang, Y., Peng, Z., Chen, X., Weng, R., Wang, J., Cai, X., Dai, W., Xiong, H.: Onecat: Decoder-only auto-regressive model for unified understanding and generation. arXiv preprint arXiv:2509.03498 (2025) 2, 3
35. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 4
36. Li, T., He, K.: Back to basics: Let denoising generative models denoise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 36115–36125 (2026) 15
37. Liang, W., Yu, L., Luo, L., Iyer, S., Dong, N., Zhou, C., Ghosh, G., Lewis, M., Yih, W.t., Zettlemoyer, L., et al.: Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. arXiv preprint arXiv:2411.04996 (2024) 3, 5
38. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025) 2, 3, 5, 12
39. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025) 2
40. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26689–26699 (2024) 12
41. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 12
42. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv preprint arXiv:2502.20321 (2025) 5
43. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. arXiv preprint arXiv:2401.08740 (2024) 4, 6
44. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7739–7751 (2025) 5, 12, 13
45. megalith: madebyollin/megalith-10m · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/madebyollin/megalith-10m (2023), [Accessed 06-11-2025] 8
46. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: Vq-vae made simple. arXiv preprint arXiv:2309.15505 (2023) 4
47. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 3
48. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025) 2, 5, 13
49. Pan, X., Singh, A., Shukla, S.N., Fan, X., Mishra, S.K., Xie, S.: Repfusion: Leveraging multimodal priors for denoising in representation space. arXiv preprint arXiv:2606.14700 (2026) 5, 15

50. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019) 10
51. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 4, 6
52. pixparse: pixparse/cc12m-wds · Datasets at Hugging Face — huggingface.co. https://huggingface.co/datasets/pixparse/cc12m-wds (2024), [Accessed 06-11-2025] 8
53. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. arXiv preprint arXiv:2412.03069 (2024) 12
54. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 13
55. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021) 13
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4, 7, 13
57. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 8
58. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025) 4
59. Singh, J., Zheng, B., Wu, Z., Zhang, R., Shechtman, E., Xie, S.: Improved baselines with representation autoencoders. arXiv preprint arXiv:2605.18324 (2026) 5
60. Song, W., Wang, Y., Song, Z., Li, Y., Sun, H., Chen, W., Zhou, Z., Xu, J., Wang, J., Yu, K.: Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. arXiv preprint arXiv:2503.14324 (2025) 5
61. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024) 4
62. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024) 2, 5, 13
63. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024) 4
64. Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164 (2024) 12
65. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024) 4

66. Tong, S., Zheng, B., Wang, Z., Tang, B., Ma, N., Brown, E., Yang, J., Fergus, R., LeCun, Y., Xie, S.: Scaling text-to-image diffusion transformers with representation autoencoders. arXiv preprint arXiv:2601.16208 (2026) 5
67. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 3, 4, 7, 10
68. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 4
69. Wang, S., Gao, Z., Zhu, C., Huang, W., Wang, L.: Pixnerd: Pixel neural field diffusion. arXiv preprint arXiv:2507.23268 (2025) 3, 4, 5, 7
70. Wang, S., Li, Z., Song, T., Li, X., Ge, T., Zheng, B., Wang, L.: Flowdcn: Exploring dcn-like architectures for fast image generation with arbitrary resolution. arXiv preprint arXiv:2410.22655 (2024) 4
71. Wang, S., Tian, Z., Huang, W., Wang, L.: Ddt: Decoupled diffusion transformer. arXiv preprint arXiv:2504.05741 (2025) 4, 5, 6, 7
72. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) 5, 12
73. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022) 4
74. Wang, Y., Lin, Z., Yang, C., Zhao, Y., Xiao, F., He, H., Zhao, Q., Ding, Z., Wang, F., Wang, S., Zhang, Y., Fan, H., Liu, X.: Representation forcing for bottleneck-free unified multimodal models. arXiv preprint arXiv:2605.31604 (2026) 5
75. Wang, Z., Bai, L., Yue, X., Ouyang, W., Zhang, Y.: Native-resolution image synthesis. arXiv preprint arXiv:2506.03131 (2025) 10
76. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025) 5, 12, 13
77. Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., Bai, X.: Liquid: Language models are scalable and unified multi-modal generators. arXiv preprint arXiv:2412.04332 (2024) 12
78. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024) 5, 12, 13
79. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025) 2, 3, 5, 7, 12, 13
80. Xu, W., Yue, X., Wang, Z., Teng, Y., Zhang, W., Liu, X., Zhou, L., Ouyang, W., Bai, L.: Exploring representation-aligned latent space for better generation. arXiv preprint arXiv:2502.00359 (2025) 4, 7
81. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 7, 10
82. Yang, J., Li, T., Fan, L., Tian, Y., Wang, Y.: Latent denoising makes good visual tokenizers. arXiv preprint arXiv:2507.15856 (2025) 4
83. Yao, J., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. arXiv preprint arXiv:2501.01423 (2025) 4, 7

84. Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025) 10, 11
85. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., et al.: Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737 (2023) 4
86. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024) 4, 5
87. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arxiv (2023) 11
88. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023) 3, 4, 7
89. Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., et al.: Pytorch fsdp: experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277 (2023) 10
90. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025) 3, 4, 5
91. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024) 2, 5
92. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 4
93. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 4