

RealGeneral: Unifying Visual Generation via Temporal In-Context Learning with Video Models

Yijing Lin Mengqi Huang Shuhan Zhuang Zhendong Mao*

University of Science and Technology of China

{lyijing, huangmq, zhuangsh}@mail.ustc.edu.cn, zdmao@ustc.edu.cn

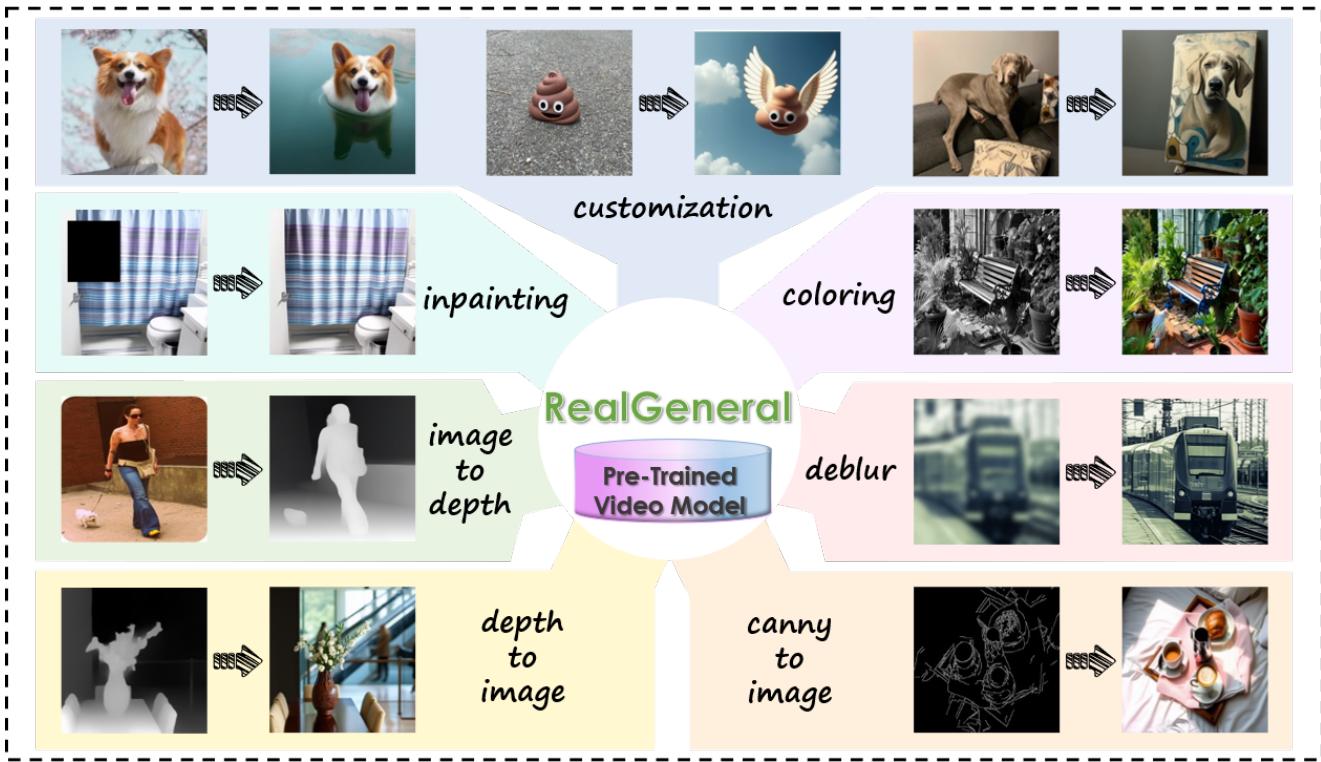


Figure 1. Results from our RealGeneral, demonstrate the ability to produce high-quality images from diverse input conditions.

Abstract

Unifying diverse image generation tasks within a single framework remains a fundamental challenge in visual generation. While large language models (LLMs) achieve unification through task-agnostic data and generation, existing visual generation models fail to meet these principles. Current approaches either rely on per-task datasets and large-scale training or adapt pre-trained image models with task-specific modifications, limiting their generalizability. In this work, we explore video models as a foundation for unified image generation, leveraging their inherent ability to model temporal correlations. We in-

introduce RealGeneral, a novel framework that reformulates image generation as a conditional frame prediction task, analogous to in-context learning in LLMs. To bridge the gap between video models and condition-image pairs, we propose (1) a Unified Conditional Embedding module for multi-modal alignment and (2) a Unified Stream DiT Block with decoupled adaptive LayerNorm and attention mask to mitigate cross-modal interference. RealGeneral demonstrates effectiveness in multiple important visual generation tasks, e.g., it achieves a 14.5% improvement in subject similarity for customized generation and a 10% enhancement in image quality for canny-to-image task. Project page: <https://lyne1.github.io/RealGeneral/>

*Corresponding author

1. Introduction

With the rapid development of generative modeling [5, 15, 16, 20, 26, 51], image generation has gained considerable attention in various application domains. Although foundation models such as Stable Diffusion [10, 32] and FLUX [22] have achieved remarkable success in unimodal text-to-image generation, processing diverse inputs (*e.g.*, text, depth maps, conditional images) and generating corresponding images within a unified framework remains a significant challenge. While large language models (LLMs) have unified text generation [1, 4, 48], an analogous unified model for the visual domain is still lacking.

Upon closer examination of the core principles underlying LLMs' success as a unified text generation architecture, we identify two fundamental pillars, *i.e.*, (1) a task-agnostic data principle during pre-training that facilitates the utilization of web-scale corpora for general linguistic knowledge acquisition, and (2) a task-agnostic generation principle that coherently unifies all kinds of textual tasks through next-token prediction. Analogously, a promising and effective architecture for unified visual generation should also exhibit (1) a **task-agnostic visual data principle** for general visual knowledge acquisition, and (2) a **task-agnostic visual generation principle** that could unify all kinds of multi-modal visual tasks (*e.g.*, customized generation, canny-to-image generation, etc.), as shown in Fig. 2.

Recent research on unified visual generation models mainly evolves along two streams, *i.e.*, (1) the training-from-scratch stream [2, 25, 33, 42, 45], which curates per-task datasets and trains unified models through *de novo optimization*, and (2) the image-model-adaptation stream [6, 54], which reformulates generation tasks as *partial image synthesis* by leveraging *pre-trained generative priors* (*e.g.*, FLUX). The first stream usually requires vast amounts of high-quality data and massive computational resources [8, 47]. For example, OmniGen[50] constructs a dataset of 0.1 billion images for all tasks and trains a unified model using 104 A800 GPUs. The second stream leverages the rich, general representations of pre-trained image generative models by fine-tuning them for specific tasks, thereby reducing computational costs. Some work [50, 53, 55] designs additional condition encoders or task-specific adapters to handle new condition types, while other work [43] revises only certain modules in the pre-trained models to adapt them to new tasks. Though much progress has been made, both of the streams fail to meet either the task-agnostic visual data principle or the task-agnostic visual generation principle. To be specific, the first stream heavily relies on the image dataset, inherently violating the task-agnostic visual data principle. Analogous to LLMs that learn by modeling relationships between sequential tokens, only video data inherently captures inter-frame relationships essential for general visual knowledge acquisition, while single im-

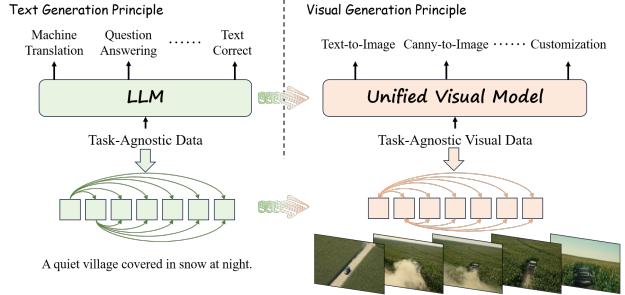


Figure 2. LLMs unify text generation through task-agnostic data and task-agnostic generation. Therefore, unified visual generation models should adhere to similar principles through visual task-agnostic data and task-agnostic visual generation principles.

ages lack this attribution. Moreover, the second stream not only depends on image data but also adopts specialized architecture for certain tasks, violating both principles.

To address the limitations of existing methods, we further explore whether large-scale pre-trained video generation models can overcome these challenges and provide a more robust framework for unified visual generation. In this study, we posit that such models, which are structurally endowed with both short-term micro-coherence and long-term macro-consistency, serve as a more effective and general foundation for this task. On the one hand, short-term micro-coherence captures fine-grained temporal dynamics, ensuring precise local details. On the other hand, long-term macro-consistency inherently preserves global structure and semantic consistency across entire sequences. Though great potential, how to effectively harness the pre-trained video models' inherent consistent generation capabilities for unified visual generation remains unexplored.

Building on this insight, we argue that video models, with their inherent capability to model temporal correlations across frames, provide a more natural foundation for unified visual generation. Inspired by in-context learning in LLMs, we reformulate image generation as a conditional frame prediction task. In this formulation, input images are treated as preceding frames in a video sequence, analogous to previous tokens in LLMs, and generate the subsequent frame as the target output. This unified formulation recasts diverse visual generation tasks as instances of conditional frame prediction, eliminating the need for task-specific architectural modifications and thereby satisfying the task-agnostic visual generation principle. Furthermore, video models are trained on task-agnostic data (*i.e.*, videos), adhering to the task-agnostic visual data principle. Therefore, the pre-trained video model only needs a small amount of corresponding task data for efficient fine-tuning, which can elicit the ability to handle various visual generation tasks, analogous to in-context learning in LLMs.

In this study, we propose a new framework termed *Re-*

alGeneral, adapting video diffusion models to image generation by proposing a minimalist framework centered on one-condition-image to one-target-image synthesis. In visual generative tasks, effectively fusing multi-modal inputs without conflating them with the generated image is challenging. To address these challenges, we introduce two core innovations. First, the *Unified Conditional Embedding* module fuses multi-modal inputs, aligning multi-modal conditions through both inter-modal conditional semantic alignment and intra-modal generative semantic distinction. Second, the *Unified Stream DiT Block*, a redesigned video model transformer block that incorporates (1) triple-branch adaLN to decouple feature modulation among text, condition frames, and target frames, eliminating the interference between the condition frame and target frame, and (2) attention mask that prevents text-to-condition interactions while preserving full condition-to-target visual attention to avoid the blending of textual and visual information.

Our main contributions are summarized as follows:

Conceptual contribution. We point out two crucial principles underlying the unification of visual generation. Furthermore, we propose a novel formulation that recasts multi-condition image generation as temporal in-context learning within video models, thereby establishing a unified framework for image generation.

Technical contribution. (1) We propose *Unified Conditional Embedding* module for aligning multi-modal inputs. (2) We propose *Unified Stream DiT Block* with decoupled adaLN and attention mask to mitigate interference between different modalities.

Experimental contribution. We validate our framework across multiple tasks. For instance, RealGeneral outperforms existing methods in the customized generation, achieving a 14.5% improvement in subject similarity compared to state-of-the-art models. In the canny-to-image task, RealGeneral enhances image quality by 10%, while in the depth-to-image task, it attains image quality comparable to existing models. Furthermore, training RealGeneral requires significantly fewer computational resources and less data than both unified models and specialized models.

2. Related work

2.1. Image Generation

Image generation is an area of rapid development that integrates concepts from natural language processing with advancements in transformer architectures. Autoregressive models utilize transformers to predict sequences of discrete codebook codes [10, 35, 40]. Meanwhile, continuous diffusion models have emerged as a powerful framework for image generation [9, 15, 37]. The integration of large-scale transformer architectures has led to the development of advanced models, such as DiT [22, 31]. Many efforts have

been made to extend the capabilities of generation models, with notable techniques, such as T2I-Adapter [28]. However, these methods are designed for specific tasks, enhancing the capabilities of generative models through architectural modifications.

2.2. Video Generation

The field of text-to-video models has witnessed remarkable progress. Early efforts to pre-train and scale Transformers for generating videos from text, such as CogVideo [18] and Phenaki [44], demonstrate significant potential. At the same time, diffusion models have recently made groundbreaking strides in video generation [3, 12, 17, 24, 39]. With the introduction of DiT, text-to-video generation has reached a new level, as highlighted by the impressive performances of Sora [29] and CogVideoX [52].

2.3. Unified Image Generation

Current research on unified image generation can be mainly categorized into two streams: training-from-scratch, and image-model-adaptation. Many works focus on the first stream [2, 8, 13, 27, 33, 42, 45]. For example, PixWizard [25] employs a flow-based DiT [31] model that gradually injects conditional and textual information during cross-attention while computing only the most relevant semantic tokens. OmniGen [50] jointly models text and images, applying causal attention to each token in the sequence and bidirectional attention within each image sequence. The image-model-adaptation stream leverages pre-trained image models to improve efficiency. For instance, ControlNet [55] uses trainable copies and zero convolution layers, enabling Stable Diffusion [10, 32] to adapt specific tasks. OminiControl [43] built on FLUX.1 [22], proposing a multi-modal attention mechanism that facilitates direct interaction between conditions and images.

Our concurrent work, UniReal [8], similarly reformulates images as discrete frames within a video generation framework as we do. However, there are three critical distinctions between RealGeneral and UniReal. First, the motivations diverge: UniReal adopts the training-from-scratch paradigm only modeling the relationship of discrete frames, whereas we propose two visual principles underlying the unification of visual generation, recasting multi-condition image generation as temporal in-context learning within video models, analogous to LLMs. Second, the design strategies are distinct. UniReal relies on complex prompt engineering, which involves multiple special tokens, contextual prompts, and hierarchical prompts, while our method doesn't need these. Third, our method is markedly more efficient: UniReal necessitates 360 million training samples and substantial computational resources. However, our method fine-tunes a pre-trained video model using LoRA with only 0.1% of that data, significantly re-

ducing the computational cost.

3. Methodology

Inspired by LLMs that unify textual generation through unlabeled large-scale text pretraining and in-context learning abilities, we argue that the video foundation model pre-trained on massive continuous visual data can similarly unify the visual generation tasks. Analogous to how LLMs autoregressively predict tokens conditioned on previous context, video models can naturally extend this concept to temporal visual frames, predicting subsequent frames from preceding ones. Specifically, given previous frames as the condition frames $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \dots\}$, video models predicts subsequent frames $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots\}$, by exploiting is temporal modeling capabilities.

In this work, we focus on a basic two-frame setting: the condition image is considered as the first frame, while the target image (accompanied by descriptive text) forms the second frame, as illustrated in Fig. 3. In this section, we first introduce our base model in Sec. 3.1 and then describe three key modules in detail. Sec. 3.2 details *Unified Conditional Embedding* (UCE) module for aligning multi-modal inputs. Sec. 3.3 explains *Separated Condition AdaLN* (SC-AdaLN) module enforcing semantic separation between condition and target frames. And Sec. 3.4 describes *Frame-Condition Decoupling* (FCD) module that employs an attention mask to prevent interactions between the condition image and text. Finally, we describe our task-specific LoRA in Sec. 3.5

3.1. Preliminaries

Our proposed model is based on CogVideoX1.5 [52], a video generation model comprising three core components: (1) a T5 text encoder [34] for text processing, (2) a 3D causal VAE for spatio-temporal compression, and (3) a transformer based on the DiT [31] architecture.

The 3D VAE encoder $\mathcal{E}(\cdot)$ compresses both spatial and temporal dimension through temporally causal convolutions, mapping an input video $\mathbf{x} \in \mathbb{R}^{(4f+1) \times 8h \times 8w \times c}$ to a latent representation $\mathbf{z}_{vision} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{(f+1) \times h \times w \times d}$, where d denotes the hidden dimension.

The transformer processes a concatenated sequence $\mathbf{z} = [\mathbf{z}_{text}, \mathbf{z}_{vision}]$, where text tokens \mathbf{z}_{text} are text embeddings. It utilizes 3D full attention for joint spatio-temporal modeling, incorporating the 3D rotary position embedding (3D-RoPE) during the computation of the query and key. Additionally, the model implements modality-specific adaptive layernorm, where video and text tokens are normalized independently and then modulated by learned scale and shift factors conditioned on their respective modalities. The denoising objective is defined in Eq. (1):

$$\mathcal{L} = \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|^2, \quad (1)$$

where ϵ denotes the noise added to the latent representations, t represents the diffusion timestep, and \mathbf{c} represents the conditioning signal.

3.2. Unified Conditional Embedding Module

In various visual generative tasks, inputs are often multi-modal (*e.g.*, text and image), while outputs are image modality. Therefore, the key to unified visual generation lies in how to effectively fuse multi-modal conditions, and avoid confusing conditional and generated images with each other. In RealGeneral, we design the UCE module to align multi-modal conditions, containing two complementary pathways, *i.e.*, the Subject-Specific Embedding layer for inter-modal conditional semantic alignment and the Condition Embedding layer for intra-modal generative semantic distinction.

As depicted in Fig. 3, 3D VAE separately encodes the condition and target images, producing latent codes $\mathbf{c}_{cond} \in \mathbb{R}^{h \times w \times d}$ and $\mathbf{x}_{target} \in \mathbb{R}^{h \times w \times d}$, respectively. Then in the UCE module, given a text embedding, we first extract subject-related embeddings $\mathbf{c}_{instance} \in \mathbb{R}^k$ using keyword matching (*e.g.*, “pot” in Fig. 3). The Subject-Specific Embedding layer then maps these embeddings to the visual latent space, denoted as $\mathbf{c}'_{instance} \in \mathbb{R}^d$. This process explicitly anchors textual subject descriptions to corresponding spatial regions in the condition image, thereby aligning textual and visual conditions. Additionally, the Condition Embedding layer applies a learnable task-specific bias $\mathbf{b}_c \in \mathbb{R}^d$ to reposition the condition embedding distribution, thereby better aligning the condition image with the target image. The enhanced condition token is obtained by combining the original embedding \mathbf{c}_{cond} with these adjustments, as Eq. (2) formulated.

$$\mathbf{c}_{cond} = \mathbf{c}_{cond} + \mathbf{c}'_{instance} + \mathbf{b}_c. \quad (2)$$

Before patchify operation, the vision sequence $\mathbf{z}_{vision} \in \mathbb{R}^{2 \times h \times w \times d}$ is formed by concatenating \mathbf{c}_{cond} and \mathbf{x}_{target} . Patchify operation in CogVideoX divides the input sequence along spatial and temporal dimensions using patch factors $p = 2$. This process compresses the temporal dimension, leading to the intermingling of the latent tokens for \mathbf{z}_{cond} and \mathbf{z}_{target} . Such mixing obscures the distinct temporal features inherent to each frame. To mitigate this, we propose a simple replication strategy, *i.e.*, we replicate the latent tokens for both \mathbf{c}_{cond} and \mathbf{x}_{target} along the temporal dimension. Finally, \mathbf{c}_{cond} , \mathbf{x}_{target} are concatenated into the input sequence for the Unified Stream DiT Block:

$$\mathbf{z} = [\mathbf{c}_{text}; \mathbf{c}_{cond}; \mathbf{x}_{target}]. \quad (3)$$

3.3. Separated Condition AdaLN Module

We introduce SC-AdaLN module, a novel mechanism to address the inherent conflict between multi-modal conditions

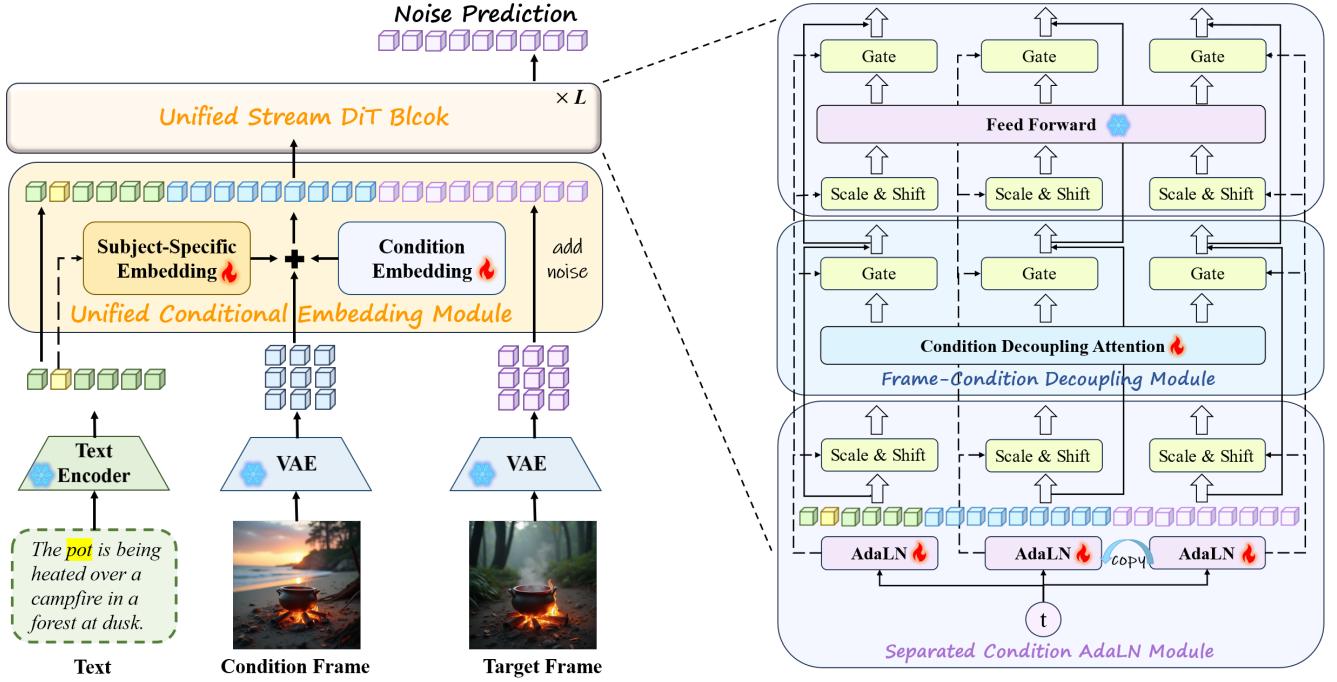


Figure 3. The overview framework of RealGeneral, recasts the image conditional generation task as a next-frame prediction. We show two frames here for simplicity. First, two images are separately encoded by VAE. The Unified Condition Embedding Module integrates textual and global task priors into the condition frame while adding noise to the target frame. Then all tokens are concatenated into a sequence entering the Unified Stream DiT Block. The Separated Condition AdaLN Module modulates text, condition, and target tokens independently via three distinct branches. The Frame-Condition Decoupling Module prevents confusion between text and condition information.

in image generation tasks. Existing video generation frameworks employ two AdaLN modules to handle text and video inputs separately. This leads to interference between the condition frame and target frame due to their shared video AdaLN parameters. Our SC-AdaLN module resolves this by decoupling this process into three independent branches, explicitly disentangling three distinct modalities:

1. **Text AdaLN** leverages the semantic information from text embeddings to provide global semantic guidance for the generation process, ensuring text-image alignment. This branch remains unchanged from the original.
2. **Condition Frame AdaLN** inherits the parameters from the original video AdaLN to preserve spatial-temporal modeling capabilities. It specializes in extracting style, structure, and local details from the condition image.
3. **Target frame AdaLN** also inherits the parameters from the original video AdaLN, which learns transition patterns between the condition and target frames.

To modulate each branch according to the current timestep t , we compute adaptive scale and shift parameters and residual gate via linear layers. Specifically, for each branch $\mathbf{k} \in \{\mathbf{c}_{text}, \mathbf{c}_{cond}, \mathbf{x}_{target}\}$, we have:

$$[\gamma_{\mathbf{k}}, \beta_{\mathbf{k}}, g_{\mathbf{k}}] = f_{\mathbf{k}}(t), \quad (4)$$

where $f_{\mathbf{k}}$ denotes a MLP that outputs the modulation coefficients, scale $\gamma_{\mathbf{k}} \in \mathbb{R}^d$, shift $\beta_{\mathbf{k}} \in \mathbb{R}^d$, and a residual gate $g_{\mathbf{k}} \in \mathbb{R}^d$ conditioned on t .

Then for each branch, the normalized feature is calculated as:

$$\text{AdaLN}_{\mathbf{k}}(\mathbf{k}; t) = \text{LN}(\mathbf{k}) \odot (1 + \gamma_{\mathbf{k}}) + \beta_{\mathbf{k}}, \quad (5)$$

where $\text{LN}(\cdot)$ denotes standard layer normalization and \odot represents element-wise multiplication. After applying AdaLN to each branch, the outputs are concatenated into a unified sequence.

3.4. Frame-Condition Decoupling Module

We propose the FCD module to mitigate multi-modal condition confusion, mainly containing the Frame-Condition Decoupling Attention (FCD Attention) which addresses the cross-modal condition interference.

In general-purpose image generation tasks, textual input primarily conveys the semantic details of the target frame, while the condition frame provides structural or stylistic guidance. In CogVideoX, text and visual tokens are jointly processed using 3D full attention, allowing unrestricted cross-modal interactions. However, this direct interaction between text tokens and condition tokens leads to the blend-

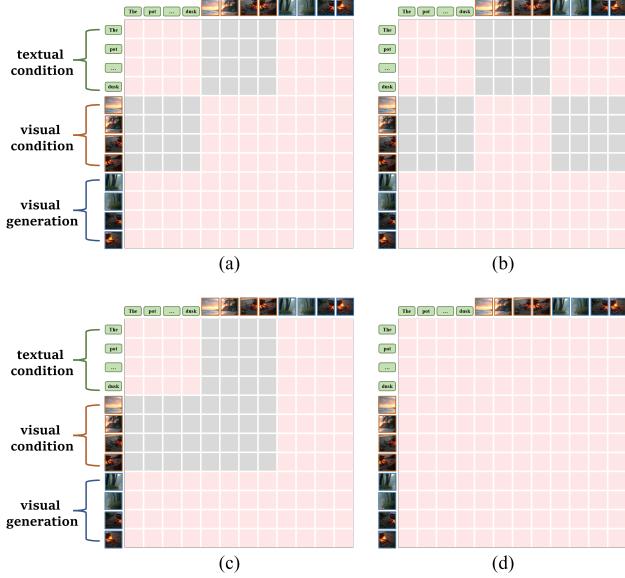


Figure 4. Various attention mask strategies. Red for interaction, and gray for blocked. (a) constrains the interaction between text and condition image. Based on (a), (b) further limits the cross-attention from the target image to the condition image, and (c) restricts the self-attention of the condition image. (d) is the original configuration without any mask.

ing of semantic and structural information, negatively affecting the quality of the generated frame.

To address this issue, we introduce the FCD Attention mechanism, which explicitly restricts interactions between text and condition tokens. Specifically, we employ a custom attention mask, M , to control these interactions precisely. The mask assigns an extremely negative value to the similarity scores between text tokens and condition tokens during attention computation, effectively decoupling the multi-modal condition interactions. This ensures that textual input influences only the target frame, while the condition frame preserves its structural properties without being overwhelmed by semantic details.

To validate the effectiveness of our design, we test various masking strategies, as illustrated in Fig. 4. Among these, the configuration shown in (a) achieves the best performance. The mask is defined as:

$$M_{ij} = \begin{cases} -\infty, & \forall (i, j) \in (\mathbf{c}_{text} \times \mathbf{c}_{cond}) \cup (\mathbf{c}_{cond} \times \mathbf{c}_{text}). \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Formally, the modified attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} + M \right) V. \quad (7)$$

3.5. Task-Specific LoRA

To efficiently adapt the billion-parameter CogVideoX, we employ LoRA [19] for fine-tuning. Specifically, for each task, we inject LoRA parameters exclusively into our proposed Separated Condition AdaLN Module and Frame-Condition Decoupling Module.

4. Experiments

4.1. Implementation Details

Task. In our experimental evaluation, we selected three tasks to evaluate our framework: subject-driven text-to-image generation, canny-to-image, and depth-to-image. The subject-driven generation is a challenging task, explored in many works [7, 11, 21, 38], evaluating the model’s ability to generate images related to a specific subject based on conditional inputs. The canny-to-image and depth-to-image tasks assess spatial understanding through edge and depth conditioning, respectively, thereby evaluating structural fidelity and three-dimensional awareness. These tasks demonstrate the model’s versatility and precision in processing diverse input modalities and generating corresponding outputs, highlighting its generalizability.

Dataset. For subject-driven text-to-image generation, we use FLUX to generate paired images of the same object with variations in scenes and poses, then filter Subject200k [43] to obtain 40K high-quality pairs, yielding a 260K two-frame video dataset. For spatially-aligned tasks (*e.g.*, canny-to-image, depth-to-image), we extract a subset from 512-2M [43], forming 310K videos, fewer than those used in previous models. All videos are resized to 512×512 . Details refer to supplementary material.

Benchmark. For customized generation, we follow previous work [7, 49], using 20 testing images in DreamBench [38] and 20 prompts. For spatially-aligned tasks, we use the COCO2017 validation set (5000 images) resized to 512×512 , with descriptions generated by ChatGPT-4o.

Metric. For subject-driven generation, we use CLIP-I, DINO [30] to calculate subject similarity and CLIP-T to measure semantic consistency. To mitigate background interference, we compute the subject similarity after segmenting the reference and generated subjects using GroundedSAM [36]. For spatially-aligned tasks, we use FID [14] and SSIM to measure image quality.

Training and Inference. Our model, based on CogVideoX1.5 and fine-tuned using LoRA with a rank and alpha of 256, comprises 892 million parameters. Training is performed over 4000 iterations across 2 days on 2 A800 GPUs, with a batch size of 44. We utilize an AdamW optimizer with a learning rate of 0.0001 and betas set to 0.9 and 0.95. During inference, we apply classifier-free guidance for the text condition, with ω values set to 6 for customization and 2 for other tasks. The denoising step is set at 50.

Method	Base-Model	CLIP-I↑	DINO↑	CLIP-T↑
BLIP-Diffusion [23]	SD 1.5	0.768	0.535	0.278
ELITE [49]	SD 1.5	0.762	0.533	0.291
SSR-Encoder [56]	SD 1.5	0.767	0.524	0.303
IP-Adapter [53]	SDXL	0.790	0.570	0.289
EMU2 [41]	SDXL	0.795	0.583	0.298
MS-Diffusion [46]	SDXL	0.772	0.540	0.312
Ours	CogVideoX	0.849	0.668	0.314

Table 1. Quantitative results with existing method on customized image generation task. We highlight the best and second-best values for each metric.

4.2. Main Results

Customized Image generation. In Fig. 5, we present qualitative comparisons with existing models, including both unified image generation models (OmniControl, OmniGen) and customization-specific models (MS-Diffusion, Emu2, SSR-Encoder). Compared to other methods, RealGeneral produces more precise details and better adherence to the prompt, underlining its effectiveness in both subject consistency and text controllability.

The quantitative results are shown in Tab. 1. Compared to existing methods, our method demonstrates significant improvement in image similarity. The image similarity is notably higher, with our approach achieving a CLIP-I score of 0.849 and a DINO score of 0.668, surpassing the second-best results by 0.054 and 0.085, respectively. At the same time, our method attains a CLIP-T score of 0.314, outperforming other methods.

Spatially-aligned tasks. In Fig. 6, we compare with existing methods. RealGeneral exhibits enhanced consistency with the provided canny edge or depth maps, generating more detailed and controllable images. Notably, because RealGeneral is based on a large-scale pre-trained video model, it essentially acquires more general visual knowledge and has a better understanding of visual relationships than those methods using image base model, as shown in Fig. 6 row 1 (*i.e.*, dog case).

The quantitative results are presented in Tab. 2. For the canny-to-image task, RealGeneral attains superior FID and SSIM scores compared to existing models. For the depth-to-image task, RealGeneral achieves a balanced performance in terms of FID and SSIM, attaining the second-best for both metrics. In addition, because the depth and canny map contain spatial information, the CLIP-T of the generated images are very close (both 0.328). We don't show this metric in the table.

4.3. Base Model Performance

We compare the image generation performance between CogVideoX1.5 and image base models (SD1.5, SDXL, FLUX), as shown in Fig. 7. The results present that

Condition	Method	Base-Model	FID↓	SSIM↑
Canny	ControlNet [55]	SD1.5	18.79	0.28
	ControlNet [55]	FLUX	98.68	0.25
	OmniGen [50]	-	22.51	0.34
	OmniControl [43]	FLUX	20.63	0.40
	Ours	CogVideoX	17.50	0.44
Depth	ControlNet [55]	SD1.5	25.12	0.24
	ControlNet [55]	FLUX	62.20	0.26
	OmniGen [50]	-	21.62	0.25
	OmniControl [43]	FLUX	27.26	0.39
	Ours	CogVideoX	23.40	0.35

Table 2. Quantitative results with existing method on canny-to-image, depth-to-image generation tasks.

Setting	CLIP-I↑	DINO↑	CLIP-T↑
w/o UCE	0.833	0.666	0.320
w/o SC-AdaLN	0.830	0.666	0.316
w/o FCD	0.827	0.665	0.318
full version	0.849	0.668	0.314

Table 3. Effect of *Unified Conditional Embedding(UCE)*, *Separated Condition AdaLN(SC-AdaLN)*, *Frame-Condition Decoupling(FCD)* modules. Removing any of the above modules causes a significant reduction in subject consistency while attaining a subtle increase in textual controllability.

CogVideoX1.5 achieves comparable image quality with SD1.5, but significantly weaker than SDXL and FLUX. Nevertheless, RealGeneral still attains better or comparable performance compared to existing methods based on pre-trained image models, demonstrating the effectiveness of the pre-training video model and the viability of unifying visual generation through task-agnostic videos.

4.4. Ablation Studies

Effect of each module. To validate the effectiveness of our proposed modules, we conduct the ablation experiments on the Unified Conditional Embedding(UCE), Separated Condition AdaLN(SC-AdaLN), and Frame-Condition Decoupling(FCD) modules, with the results shown in Tab. 3. Removing UCE module results in a significant decrease in CLIP-I ($0.849 \rightarrow 0.833$), indicating its critical role in aligning multi-modal semantics. Similarly, the removal of SC-AdaLN or FCD module leads to reductions in CLIP-I by 0.019 and 0.022, respectively, showcasing the importance of decoupling multi-modal conditions. Although these modifications cause a slight decrease in CLIP-T, the improvement in topic similarity remains noteworthy.

Impact of various attention mask strategies. To assess the impact of different attention mask strategies on model performance, we conduct ablation experiments as shown in Tab. 4. Among the several attention mask strategies de-

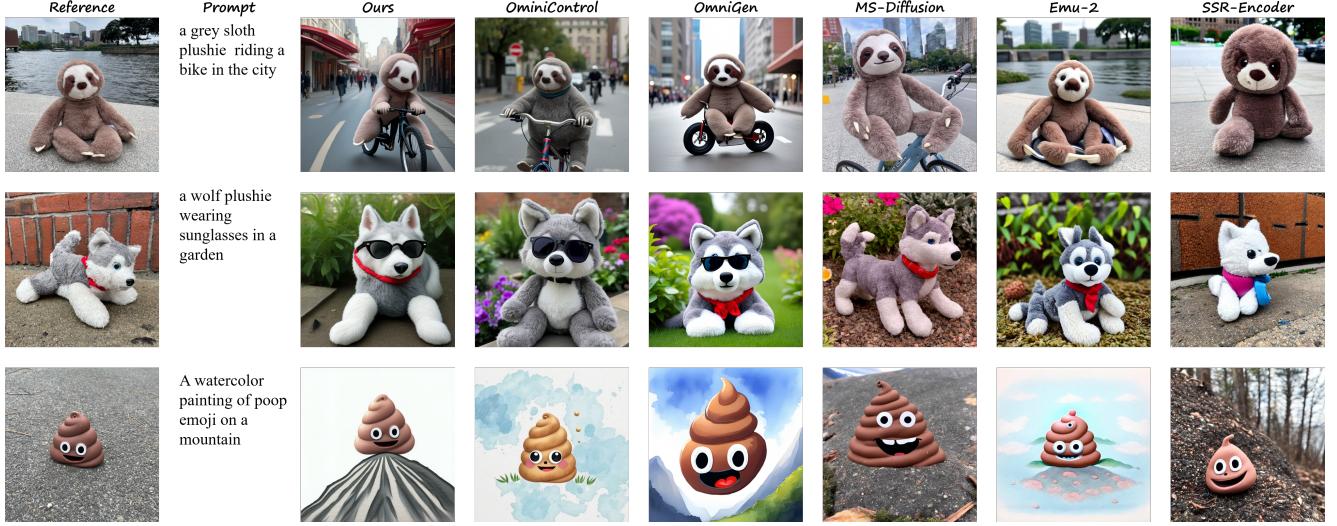


Figure 5. Qualitative comparison with existing methods, consisting of unified image generation models and customization-specific models. The results demonstrate that RealGeneral exhibited superior performance in terms of subject consistency and text consistency.



Figure 6. Qualitative comparison with existing methods on canny-to-image task. Our method is more consistent with the given depth or canny map than other methods.



Figure 7. Comparison of image generation results between CogVideoX1.5 and other image base models.

picted in Fig. 4, Mask A, which blocks the interaction between the textual input and the condition image, achieves the highest scores in subject similarity with only a minimal reduction in textual similarity. In contrast, both Mask B and the no-mask result in lower subject consistency, and Mask C significantly impairs the model’s performance in both subject consistency and textual controllability by restricting the condition image’s self-attention. The qualitative results are

Setting	CLIP-I↑	DINO↑	CLIP-T↑
Mask B	0.827	0.662	0.321
Mask C	0.779	0.650	0.312
no mask	0.827	0.665	0.318
Mask A	0.849	0.668	0.314

Table 4. Ablation study on various attention masks.

shown in supplementary material.

5. Conclusion & Future Work

In this paper, we propose RealGeneral, a novel framework that bridges the gap between video generation models and unified visual synthesis tasks. we unlock the inherent potential of pre-trained video models for diverse image generation tasks, by reformulating conditional image generation as temporal in-context learning through temporal modeling. This is achieved through three core modules: (1) UCE module that aligns multi-modal semantics, (2) SC-AdaLN that disentangles feature modulation across text, conditional frames, and target frames, (3) FCD module that prevents conditional semantic interference. Together, these components form a solution for unified image generation tasks.

Future Work. Future research will explore several directions to further advance RealGeneral. First, we plan to validate the framework on video models with higher generation quality and larger pre-training to assess its generalizability. Second, we will focus on developing enhanced video foundation models that generate higher-quality image synthesis. Finally, we aim to develop a fully integrated, universal framework instead of task-specific LoRA fine-tuning.

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
- [2] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024. 2, 3
- [3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 3
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2
- [6] Haodong Chen, Lan Wang, Harry Yang, and Ser-Nam Lim. Omnicreator: Self-supervised unified generation with universal editing. *arXiv preprint arXiv:2412.02114*, 2024. 2
- [7] Nan Chen, Mengqi Huang, Zhuowei Chen, Yang Zheng, Lei Zhang, and Zhendong Mao. Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization. *arXiv preprint arXiv:2409.05606*, 2024. 6
- [8] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. *arXiv preprint arXiv:2412.07774*, 2024. 2, 3
- [9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3
- [10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3
- [11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 6
- [12] 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. 3
- [13] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024. 3
- [14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3
- [16] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2
- [17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3
- [18] 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. 3
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6
- [20] Mengqi Huang, Zhendong Mao, Zhuwei 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. 2
- [21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 6
- [22] Black Forest Labs. Flux: Official inference repository for flux.1 models. <https://github.com/black-forest-labs/flux>, 2024. Accessed: 2024-11-12. 2, 3
- [23] 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. 7
- [24] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han, and Jun-Yan Zhu. Efficient spatially sparse inference for conditional gans and diffusion models. *Advances in neural information processing systems*, 35:28858–28873, 2022. 3
- [25] Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Junlin Xie, Yu Qiao, Peng Gao, and Hongsheng Li. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. *arXiv preprint arXiv:2409.15278*, 2024. 2, 3

- [26] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. *arXiv preprint arXiv:2501.13554*, 2025. 2
- [27] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023. 3
- [28] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 3
- [29] OpenAI. Sora. <https://openai.com/index/sora/>, 2024. 3
- [30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6
- [31] 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. 3, 4
- [32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3
- [33] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 2, 3
- [34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 4
- [35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 3
- [36] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 6
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 6
- [39] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3
- [40] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 3
- [41] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 7
- [42] Zeyi Sun, Ziyang Chu, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. X-prompt: Towards universal in-context image generation in auto-regressive vision language foundation models. *arXiv preprint arXiv:2412.01824*, 2024. 2, 3
- [43] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 2, 3, 6, 7
- [44] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3
- [45] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2, 3
- [46] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *The Thirteenth International Conference on Learning Representations*, 2025. 7
- [47] Zhaoqing Wang, Xiaobo Xia, Runnan Chen, Dongdong Yu, Changhu Wang, Mingming Gong, and Tongliang Liu. Lavindit: Large vision diffusion transformer. *arXiv preprint arXiv:2411.11505*, 2024. 2
- [48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2
- [49] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 6, 7

- [50] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuteng Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. [2](#), [3](#), [7](#)
- [51] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking “text” out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8682–8692, 2024. [2](#)
- [52] 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. [3](#), [4](#)
- [53] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. [2](#), [7](#)
- [54] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueling Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024. [2](#)
- [55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. [2](#), [3](#), [7](#)
- [56] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024. [7](#)