Dimension-Reduction Attack! Video Generative Models are Experts on Controllable Image Synthesis

Hengyuan Cao
Zhejiang University

Yutong Feng[†]
Kunbyte AI

Biao Gong
Ant Group

Yijing Tian
Hangzhou Normal University

Yunhong Lu
Zhejiang University

Chuang Liu
Hangzhou Normal University

Bin Wang
Kunbyte AI

{caohy, yunhonglu}@zju.edu.cn tianyijing2002@163.com liuchuang@hznu.edu.cn
{fengyutong.fyt, a.biao.gong, binwang393}@gmail.com

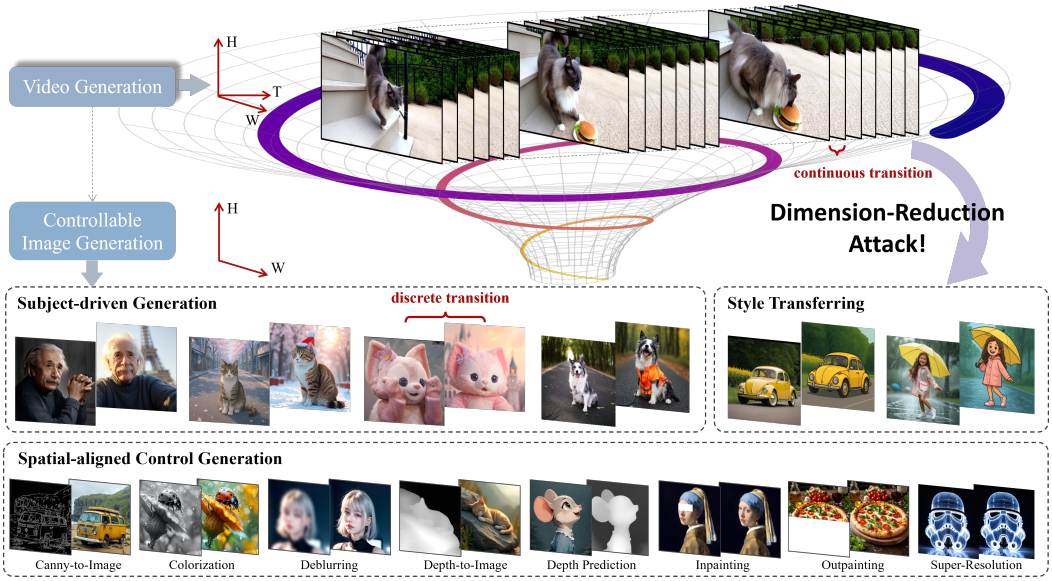


Figure 1: This paper leverages high-level prior of video generative models to unify controllable image generation in low-level. Bottom results show various types of task supported by `DRA-Ctrl`.

Abstract

Video generative models can be regarded as world simulators due to their ability to capture dynamic, continuous changes inherent in real-world environments. These models integrate high-dimensional information across visual, temporal, spatial, and causal dimensions, enabling predictions of subjects in various status. A natural and valuable research direction is to explore whether a fully trained video generative model in high-dimensional space can effectively support lower-dimensional tasks such as controllable image generation. In this work, we propose a paradigm for video-to-image knowledge compression and task adaptation, termed *Dimension-Reduction Attack* (`DRA-Ctrl`), which utilizes the strengths of video models, including long-range context modeling and flatten full-attention, to perform various generation tasks. Specially, to address the challenging gap between

[†]Project leader.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

continuous video frames and discrete image generation, we introduce a mixup-based transition strategy that ensures smooth adaptation. Moreover, we redesign the attention structure with a tailored masking mechanism to better align text prompts with image-level control. Experiments across diverse image generation tasks, such as subject-driven and spatially conditioned generation, show that repurposed video models outperform those trained directly on images. These results highlight the untapped potential of large-scale video generators for broader visual applications. `DRA-Ctrl` provides new insights into reusing resource-intensive video models and lays foundation for future unified generative models across visual modalities. The project page is `https://dra-ctrl-2025.github.io/DRA-Ctrl/`.

1 Introduction

Recent advances in text-to-image (T2I) generative models [38, 34, 9, 23] have significantly improved the quality of image synthesis from natural language prompts. To enhance controllability, researchers have introduced auxiliary conditions into the context of generation [54, 24, 49, 32, 60, 10, 18], such as subject reference images, edge maps and depth cues. This has given rise to the paradigm of *controllable image generation*, where both textual and visual conditions collaboratively guide the synthesis process. While early methods relied on additional image adapters or cross-attention mechanisms [54, 6, 15, 42, 44], recent approaches leverage *full-attention* architectures [61, 22, 50, 43, 51, 7, 25, 12, 29, 47] that treat all input tokens as a unified sequence. However, these models are all built upon image generative models, thus remain limited by the static nature of image data, which lacks the continuous temporal and causal structures transformation present in the real world.

Video generative models [2, 53, 20, 45], in contrast, are trained to predict sequences of frames with rich spatiotemporal dependencies. The prior knowledge learned by these models incorporates long-range context, consistent object transitions, non-rigid transformation and high-level scene dynamics. These capabilities align closely with the goals of controllable image generation. This observation inspires a new direction, *i.e.,* repurposing pretrained video models to support image-level tasks by transferring their high-dimensional knowledge into a lower-dimensional setting. This work dives into this idea and presents a framework termed `DRA-Ctrl` that efficiently adapts video generators for diverse controllable image generation scenarios.

However, directly adapting video generative models to controllable image generation presents non-trivial challenges. A naive baseline would be to gather the condition image and target image into the frame sequence of video generators. The key hindrance confronted here is that the video data inherently consists of temporally continuous frames with smooth transitions, while the condition-target image pairs represent a discrete, abrupt change between two states. In detail, we investigate to adapt two variants of video generative models treating the image pairs as two-framed video. For image-to-video (I2V) model consuming the condition image as the first frame, it suffers to over-constrain the output to mimic the condition image. While for text-to-video (T2V) model, it is inevitable to inject the condition image as non-noisy frame tokens into the sequence. Thus the model takes much efforts to readapt the new paradigm, and tends to forget its pre-training knowledge with suboptimal performance. These baseline solutions expose the fundamental discrepancy between the continuous dynamics learned by video models and the discrete transition required by controllable image generation. Therefore, it is essential for `DRA-Ctrl` to conduct stable transferring when repurposing the video models without forgetting their high-dimensional capabilities.

To address these challenges in `DRA-Ctrl`, we propose a *mixup-based transition* strategy, inspired by the mixup [57] principle in representation learning, serving as a bridge connecting the diverse intermediate gaps in videos and images. The core idea is to treat the condition and target images as boundary frames of a synthetic shot transition sequence, with intermediate frames generated using a temporal position-aware mixup. Each intermediate frame is weighted by its relative position between the two endpoints, enabling smooth interpolation while preserving key visual characteristics. We implement the mixup transition with the I2V model. When integrating with these augmented frames, the constraint from condition to target images is significantly relaxed, making it easier to adapt to discrete image generation. Despite this, real video transitions generally require dozens of intermediate frames, resulting in dramatically increased computation cost. To mitigate this, we introduce *Frame-Skip Position Embedding*, a positional encoding scheme that expands temporal intervals in the latent space, allowing large image transformations with only a few frames. Additionally, to distinguish

complex combination of subjects and environments in multiple images, we adapt the condition and target prompts into the full-attention mechanism together with a masking strategy.

We evaluate DRA-Ctrl on a wide range of controllable image generation tasks, including subject-driven image synthesis, spatially aligned condition generation (*e.g.,* canny-to-image translation, colorization, deblurring, depth-based generation and depth prediction), masking image generation (inpainting and outpainting) and style transferring. Our experiments demonstrate that video generative models can be effectively re-purposed for these tasks, consistently outperforming methods built upon image generative models. This surprising effectiveness highlights a compelling "*Dimension-Reduction Attack*", where high-dimensional video priors offer enhanced control when adapted to lower-dimensional image tasks, encouraging more efforts to further investigate the extending capability of video generative models.

2 Related Works

Subject-driven Image Generation. Subject-driven image generation with diffusion models typically follows two paradigms: tuning-based and tuning-free methods. Tuning-based methods [40, 11, 16, 21] achieve strong identity consistency but require per-subject fine-tuning, limiting scalability and introducing non-trival computational overhead. Tuning-free methods instead enhance generalization through training on large-scale datasets, eliminating inference-time tuning. Early works [54, 24, 49, 32, 60, 18] extract subject information from reference images using an image encoder, and inject these features into the generation process via cross-attention mechanisms. Then, Hu et al. [15] propose using a ReferenceNet which is architecturally identical to the denoising UNet as the image feature extractor, providing detailed and accurate control information for controllable generation. Later advancements in tuning-free methods leverage the model's inherent in-context learning capabilities [17], treating the model itself as an image feature extractor to provide subject-specific information for generation. Zeng et al. [56] proposes to model the joint distribution of multiple text-image pairs sharing the same subject, investigating in-context learning within UNet-based diffusion models for subject-driven image generation. With the introduction of DiT architectures [33], recent works [61, 22, 50, 43, 51, 7, 25] have explored the full-attention mechanism, where reference images and generated images jointly participate in self-attention, to facilitate subject feature extraction and enable in-context learning for subject-driven generation. We propose leveraging video diffusion models' inherent frame-level full-attention mechanism for subject-driven generation.

Spatially-aligned Image Generation. Spatially-aligned control signals for fine-grained image generation have emerged as a critical research direction. Early conditional Generative Adversarial Networks (GANs) [19, 63] and transformers [4] achieve image-to-image translation by learning the mapping from conditional images to target images. Recent diffusion models enable tighter integration of such controls. SDEdit [30] guides generation process by first adding noise to stroke paintings and then denoising them. In contrast, T2I-Adapter [31] trains an adapter network to enable more diverse and precise control signals. ControlNet [58] reuses the encoding layers of pre-trained diffusion models as a backbone for learning control signals. UniControl [36] further advances this direction by integrating multiple tasks within a unified framework via a task-aware HyperNet, demonstrating zero-shot capabilities on unseen tasks and combined tasks. Subsequent works [7, 51, 43, 25, 12, 29, 47] have unified subject-driven and spatially-aligned image generation within one framework that maps control images to target outputs, which DRA-Ctrl also follows.

Image Generation with Video Models. While existing works employ video generative models for *image editing* (requiring pixel-aligned partial modifications) that are methodologically naive, our framework targets *controllable image generation* that enables comprehensive transformations — including background replacement, subject pose/state alteration, and holistic content regeneration. FramePainter [59] injects interactive editing signals extracted by the control encoder into the generation process via cross-attention mechanisms and synthesizes a two-frame video where the first frame reconstruct the condition image and the second one produces the edited output. ObjectMover [55] addresses the object relocation task by fine-tuning a video generative model through frame-wise concatenation of condition images with various control signals. Rotstein et al. [39] proposes a direct I2V approach for image editing, where condition images and Vision Language Model (VLM)-processed prompts are jointly fed into the model, with edited results obtained through a specialized frame selection strategy. While Lin et al. [27] and Chen et al. [7] similarly employ video models for controllable image generation or editing tasks, primarily motivated by their ability

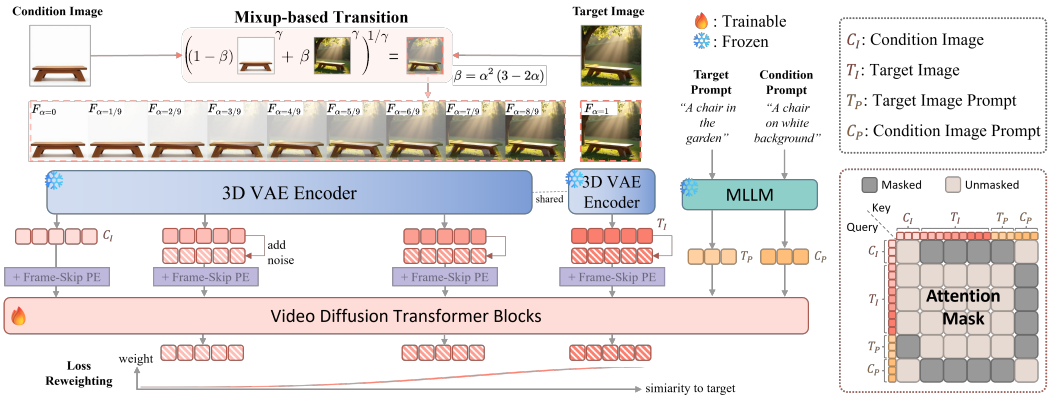


Figure 2: The **training framework** of `DRA-Ctrl`. We propose a mixup-based transition strategy to construction shot transition videos to adapt the video model for abrupt image changes, with FSPE strategically reducing transitional frames. The loss function is adaptively reweighted according to the proportion of target image in the token sequence. Besides, to align text prompts with image-level control, we design an attention masking mechanism.

to perform full attention in the temporal dimension, our work further introduces strategies like mixup to better exploit the rich priors inherent in video models.

3 Method

Given that video generative models' inherent **temporal full-attention** and **rich dynamics priors**, we argue they can be efficiently re-purposed for controllable image generation tasks. To successfully adapt smooth-transition-capable video generative models for handling abrupt and discontinuous image transitions, we propose multiple strategies, as shown in Figure 2. Specifically, in Section 3.1, we introduce our foundational model, HunyuanVideo-I2V, detailing its architecture and objective function; in Section 3.2, we present our mixup-based shot transition strategy that construct a shot transition video with condition and target images; in Section 3.3, we propose a new position embedding method that reduces the required number of transition frames; in Section 3.4, we describe an attention masking strategy to properly guide information interaction.

3.1 Preliminaries

Our method builds upon HunyuanVideo-I2V [20], which consists of three key components: (1) a causal 3DVAE that compresses videos in both spatial and temporal dimensions, (2) a text encoder built upon a Multimodal Large Language Model (MLLM), which processes not only textual information but also partial conditioning image features, (3) a transformer employing a unified full-attention mechanism to jointly process image and text signals.

The 3DVAE maps a video sequence $\mathbf{x} \in \mathbb{R}^{(4T+1)\times 3\times 16H\times 16W}$ into a compact latent representation $\mathbf{y} \in \mathbb{R}^{(T+1)\times 16\times 2H\times 2W}$, which is subsequently patchified and unfolded to yield visual tokens $\mathbf{z}_{visual}$ of length $(T+1)\times H\times W$. Meanwhile, the textual tokens $\mathbf{z}_{textual}$ are obtained by processing target prompt T_P and condition image C_I through the MLLM. Then a concatenated sequence $\mathbf{z} = [\mathbf{z}_{visual}, \mathbf{z}_{textual}]$ is fed into the transformer, where a unified full-attention mechanism is applied to effectively fuse information across both modalities. To enhance the model's ability to capture positional relationships, 3D Rotary Position Embedding (RoPE) [41] is introduced in each transformer block. To achieve I2V generation, HunyuanVideo-I2V employs a token replacement technique, where the visual tokens of the first frame are replaced with the condition image tokens. In addition, CLIP-Large [37] text features and the diffusion timestep t are adopted as global guidance signals and incorporated into the transformer. The objective function follows flow matching [28]:

$$\mathcal{L} = \|v_\theta(\mathbf{y}_t, t, C_I, T_P) - (\epsilon - \mathbf{y})\|^2, \tag{1}$$

where ϵ denotes Gaussian noise, $\mathbf{y}_t = (1-t)\mathbf{y} + t\epsilon$, and v and θ stand for the neural network and its corresponding parameters respectively.

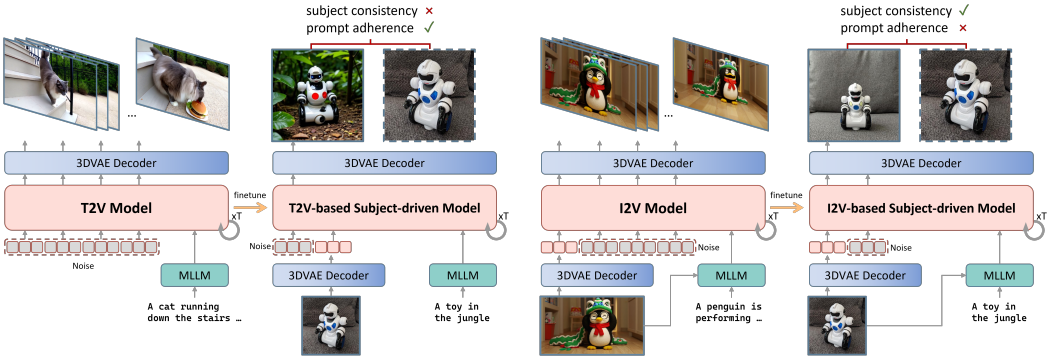


Figure 3: The inference process of T2V/I2V models and their finetuned subject-driven image generation models. By treating the condition and target images directly as a two-frame video and fine-tuning T2V/I2V models accordingly, the corresponding T2V/I2V baselines can be obtained.

3.2 Mixup-based Shot Transition

The simplest approach for controllable image generation using video generative models is to treat condition and target images as a two-frame video. During training, the condition image remains noiseless and excluded from loss calculation, while the target image is noise-corrupted and included in loss calculation. During inference, the condition image maintains noiseless to provide complete control signals. Empirical tests with HunyuanVideo-T2V/I2V on subject-driven generation task, as shown in Figure 3 and Table 3, demonstrate that neither model meets the requirements for subject-driven generation: the T2V model lacks subject consistency, while the I2V model over-preserves similarity to the condition image and exhibits poor prompt adherence. The observed results are expected because T2V model does not enforce consistency as strictly as I2V model, while the I2V model's strong inter-frame consistency preservation limits prompts' controllability.

To address these limitations, we draw inspiration from cinematic shot transitions by treating condition and target images as storyboard endpoints. Then, we fine-tune the I2V model to generate transition frames and target image according to condition image. This approach maintains consistency and enhances controllability through smooth visual transitions. Specifically, we observe that certain I2V models [20, 45] can naturally produce fade-in-fade-out transitions similar to those in PowerPoint presentations.Therefore, we propose constructing transition frames F_α with condition image $F_{\alpha=0}$ and target image $F_{\alpha=1}$ by interpolation, $F_\alpha = \left((1-\beta)\, F_{\alpha=0}^{\gamma} + \beta F_{\alpha=1}^{\gamma}\right)^{1/\gamma}$, $\beta = \alpha^2\,(3-2\alpha)$, where $\alpha \in [0,1]$ and γ is set to 2.2 ensure smooth inter-frame transitions. During training, we keep the condition image $F_{\alpha=0}$ noise-free and exclude it from loss calculation, while applying noise and including $F_{0<\alpha\leq 1}$ in the loss calculation. The contribution weight of each latent frame in the loss is determined by its proportional content from the target image, yielding the final loss function:

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{K+1}\sum_{k=0}^{K} w(k)\|v_\theta(\mathbf{y}_t^k, t, C_I, C_P, T_P) - (\epsilon - \mathbf{y})\|^2, \\
\mathbf{y}_t^k &= \begin{cases} (1-t)\cdot Encode(F_{0\leq\alpha<1}, k) + t\epsilon, & \text{if } k = 0,1,\ldots,K-1, \\ (1-t)\cdot Encode(F_{\alpha=1}, -1) + t\epsilon, & \text{if } k = K, \end{cases} \\
w(k) &= \frac{1}{4}\sum_{i=1}^{4}\left(\left(\frac{4k+i}{4K+1}\right)^2\left(3-2\frac{4k+i}{4K+1}\right)\right)^2,
\end{aligned}
\tag{2}
$$

where $Encode(\cdot, k)$ is the encoder of the 3DVAE, which maps $4T+1$ frames in pixel space to $T+1$ latent representations and returns the $(k+2)$-th latent representation, C_P is the prompt of the condition image. We encode the target image separately to ensure the independence of the corresponding latent representation during inference. During inference, the condition image's latent representation is concatenated with $K+1$ Gaussian noise in latent space and perform progressive denoising while keeping the condition image's latent representation unchanged throughout the process, ultimately decoding the last frame of the denoised latent representations through the decoder $Decode(\cdot)$ of the 3DVAE to obtain the final generated result $\hat{F}_{\alpha=1} = Decode(\mathbf{y}^K)$.

3.3 Frame Skip Position Embedding

Achieving smooth shot transition often requires dozens or even hundreds of frames. Since we only aim to obtain the final frame, inserting so many transition frames between condition and target images would severely degrade the efficiency of both training and inference. In HunyuanVideo, the model incorporates both temporal and spatial information (n, i, j) into tokens through RoPE, where $n = 0, 1, \cdots, T$ represents the latent frame index of the tokens in temporal dimension and $i = 0, 1, \cdots, H-1$ and $j = 0, 1, \cdots, W-1$ denote the height and width coordinates of the tokens in spatial dimensions, respectively. To achieve long-term effects with minimal latent frames, we enhance RoPE by incorporating skip intervals along the temporal dimension, called Frame Skip Position Embedding (FSPE), $(n', i', j') = (n \times \delta, i, j)$, where δ represents the skip interval. This approach constructs a long-term sparse representation of latent frames using minimal latent frames, significantly reducing computational overhead.

3.4 Attention Masking Strategy

Due to the absence of textual descriptions for shot transition videos, we jointly input the prompts from both condition and target image into the network on subject-driven generation task. This approach enables the model to acquire all textual information corresponding to the shot-transition videos. However, in this way, there are four distinct token sequences during full-attention computation, i.e., condition image tokens C_I, generated frame tokens T_I, target image prompt tokens T_P, and condition image prompt tokens C_P. To prevent unintended information blending across these token sequences, we design an attention masking strategy as illustrated in Figure 2. Specifically, our designed attention mask assigns a extremely negative value to similarity scores between incompatible token sequences (e.g., condition image tokens and target image prompt tokens) to effectively block unintended interactions while maintaining necessary information flows,

$$M_{pq} = \begin{cases} -\infty, & \text{if } (p,q) \in (C_I \times T_I) \cup (T_I \times C_P) \cup (T_P \times C_I) \cup (T_P \times C_P) \cup (C_P \times T_I), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Furthermore, during inference, we enhance the differentiation between T_P and C_P influences by augmenting the attention mask region corresponding to $(T_I \times T_P)$ with an offset of ω times its absolute mean value, where we set $\omega = 0.6$.

4 Experiments

4.1 Experimental Setup

Tasks. We extensively evaluate the effectiveness of our method across multiple tasks, including spatially-aligned generation, subject-driven generation and style tranferring. For spatially-aligned image generation, we specifically design five distinct sub-tasks: canny-to-image generation, depth-to-image generation, image colorization, image deblurring and image in/out-painting.

Training. For spatially-aligned image generation, we adopt a subset of the Text-to-Image-2M dataset [64] for training, consisting of around 160K samples, where the condition images are extracted from the corresponding ground-truth images. The models are trained with a batch size of 8 and gradient accumulation over 2 steps, resulting in an effective batch size of 16. We employ the AdamW optimizer and conduct training on 2 NVIDIA H800 GPUs (80GB memory each). For subject-driven image generation, we utilize the high-quality subset of the Subjects200K dataset [43], comprising approximately 110K image pairs for training. This model is trained using 4 NVIDIA H800 GPUs.

Benchmarks. For spatially-aligned generation, we employ the COCO2017 validation dataset [26] comprising 5,000 images resized to 512×512 resolution as the test set, where the corresponding prompts are randomly selected from multiple candidate captions associated with each image. For subject-driven generation, we evaluate our method on DreamBench [40] by generating images for 25 text prompts per subject, using one reference image for each of the 30 subjects in the benchmark.

Metrics. For spatially-aligned generation, we evaluate methods in terms of controllability and generation quality. Controllability is assessed by the similarity of the extracted condition images from generated and ground-truth image. Specifically, we employ the F1 score for canny-to-image

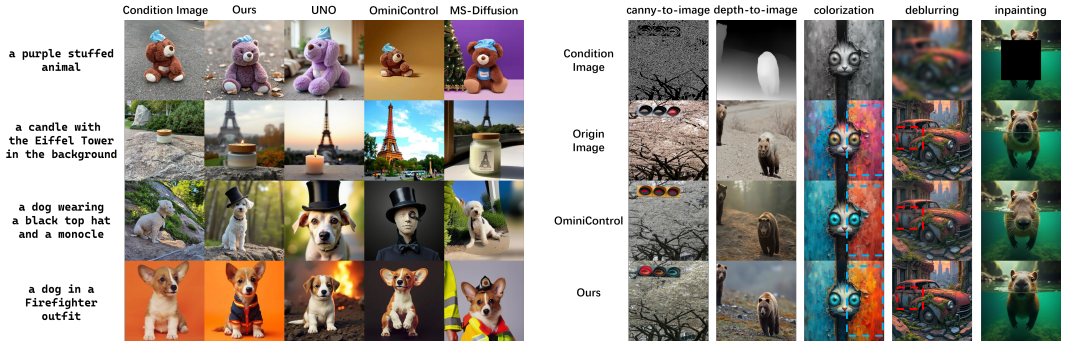
Table 1: Quantitative results on COCO2017 validation set. The best results are in **bold**.

Condition	Model	Method	Controllability F1↑/MSE↓	General Quality FID↓	SSIM↑
Canny	SD1.5 [38]	ControlNet [58]	0.34	18.74	0.35
		T2I-Adapter [31]	0.22	20.06	0.35
		Uni-ControlNet [62]	0.20	17.38	–
	FLUX.1 [23]	ControlNet	0.21	98.68	0.25
		OminiControl [43]	0.38	20.63	**0.40**
		EasyControl [61]	0.31	**16.07**	–
	HunyuanVideo-I2V [20]	DRA-Ctrl	**0.42**	19.44	0.38
Depth	SD1.5	ControlNet	923	23.02	0.34
		T2I-Adapter	1560	24.72	0.27
		Uni-ControlNet	1685	21.79	–
	FLUX.1	ControlNet	2958	62.20	0.26
		OminiControl	903	27.26	**0.39**
		EasyControl	1092	**20.39**	–
	HunyuanVideo-I2V	DRA-Ctrl	**76**	20.83	0.33
Deblur	FLUX.1	ControlNet	572	30.38	0.74
		OminiControl	132	11.49	**0.87**
	HunyuanVideo-I2V	DRA-Ctrl	**11**	**9.08**	0.64
Colorization	FLUX.1	ControlNet	351	16.27	0.64
		OminiControl	**24**	10.23	0.73
	HunyuanVideo-I2V	DRA-Ctrl	30	**8.39**	**0.85**
Mask	SD1.5	ControlNet	7588	13.14	0.40
	FLUX.1	OminiControl	6248	15.66	0.48
	HunyuanVideo-I2V	DRA-Ctrl	**16**	**9.87**	**0.59**

task and use Mean Squared Error (MSE) for other tasks. Generation quality is quantified using Fréchet Inception Distance (FID) [13] and Structural Similarity Index Measure (SSIM) [48] between generated and ground-truth images. For subject-driven generation, we evaluate methods by standard automatic metrics and a Vison-Language (VL) Model. We measure subject consistency by DINO and CLIP-I scores, which compute the cosine similarity between the condition image and the generated image in DINO [3] and CLIP [37] embedding spaces. Prompt adherence is quantified by the cosine similarity between the CLIP embeddings of the prompt and the generated image, referred to as CLIP-T score. However, these metrics have inherent limitations: DINO and CLIP-I measure global image similarity rather than directly evaluating subject consistency, while CLIP-T struggles with fine-grained semantic alignment and other challenges [37]. To address this, we propose VL score, a novel metric based on QWen2.5-VL [1], which evaluates generated images for subject consistency and prompt adherence via tailored prompts. The VL model outputs discrete scores (0-4) per dimension, with the final score computed as their average.

4.2 Spatially-aligned Image Generation Results

To validate DRA-Ctrl's effectiveness for spatially-aligned generation tasks, we conduct comprehensive comparisons with multiple competitive approaches. As shown in Figure 4b, our method demonstrates superior performance in several aspects: compared to OminiControl, our approach generates more realistic traffic light images for canny-to-image; produces images with more vivid details for depth-to-image; achieves richer color variations in the blue-boxed regions for colorization; better preserves original image details in red-boxed areas for deblurring; and creates more authentic results for inpainting. These qualitative comparisons consistently highlight our method's advantages in maintaining spatial alignment while generating high-quality images across diverse generation scenarios. Quantitative results presented in Table 1 further demonstrate the superiority of DRA-Ctrl. DRA-Ctrl achieves significant advantages in controllability, attaining the best results across all tasks except colorization, while maintaining highly competitive performance in general quality.



(a) Subject-driven image generation. (b) Spatially-aligned image generation.

Figure 4: Qualitative results comparing different methods.

Table 2: Quantitative results on DreamBench. The **best** and **second best** values of each metric are highlighted.

Method	VL Score↑	DINO↑	CLIP-I↑	CLIP-T↑
Oracle	–	0.774	0.885	–
Textual Inversion [11]	–	0.569	0.780	0.255
DreamBooth [40]	–	0.668	0.803	0.305
BLIP-Diffusion [24]	–	0.670	0.805	0.302
ELITE [49]	–	0.647	0.772	0.296
Re-Imagen [5]	–	0.600	0.740	0.270
BootPIG [35]	–	0.674	0.797	0.311
SSR-Encoder [60]	–	0.612	0.821	0.308
OmniGen [51]	–	0.693	0.801	**0.315**
OminiControl [43]	2.21	0.559	0.765	0.310
FLUX.1 IP-Adapter [54]	–	0.582	**0.820**	0.288
MS-Diffusion [46]	1.94	0.655	0.782	0.307
UniReal [7]	–	**0.702**	0.806	**0.326**
UNO [50]	**2.43**	0.657	0.786	**0.315**
DRA-Ctrl	**2.56**	**0.722**	**0.825**	0.302

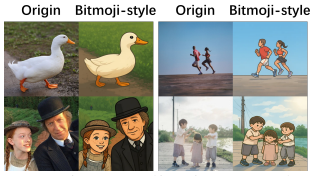


Figure 5: Qualitative results of DRA-Ctrl on style transferring.

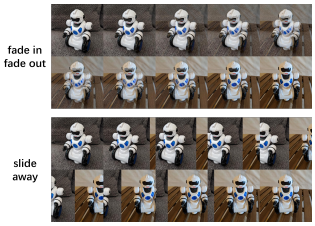


Figure 6: Different mixup-based shot transition types.

4.3 Subject-driven Image Generation Results

To validate the effectiveness of DRA-Ctrl for subject-driven generation, we conduct comprehensive comparisons with multiple state-of-the-art approaches. Qualitative results are presented in Figure 4a, where our method demonstrates superior subject consistency. As shown in the third row, our approach generates a dog that even preserves details like the neck tag, while competing methods exhibit inconsistent breeds or fail to generate the subject altogether. The quantitative results are presented in Table 2, where we compare various tuning-based and tuning-free approaches. Under all comparison methods, our approach achieves the highest VL Score (2.56), DINO (0.722), and CLIP-I (0.825), along with a competitive CLIP-T score of 0.302.

4.4 Style Transfer

We employ GPT-4o to generate 100 original-to-Bitmoji-style image pairs, which are subsequently used to fine-tune our subject-driven model for achieving style transfer effects. The results are demonstrated in Figure 5, where our model successfully captures the distinctive aesthetic characteristics of Bitmoji-style animation while preserving the original content's structural integrity.

4.5 Ablation Studies

To validate the effectiveness of our proposed strategies, we conduct comprehensive ablation studies on our method from multiple perspectives, including comparisons with T2V/I2V baselines, analysis of different shot transition types, ablation on the number of transition frames, and module ablation.

Table 3: Comparison with baselines.

	VL↑	DINO↑	CLIP-I↑	CLIP-T↑
Oracle	–	0.774	0.885	–
T2V baseline	2.01	0.658	0.787	**0.306**
I2V baseline	2.34	**0.803**	**0.874**	0.291
DRA-Ctrl	**2.44**	0.715	0.821	0.298

Table 4: Ablation on transition types.

	VL↑	DINO↑	CLIP-I↑	CLIP-T↑
slide away	2.19	0.708	0.822	0.292
fade in fade out	**2.42**	**0.742**	**0.834**	**0.295**

Table 5: Ablation on frame numbers.

number of transition frames	VL↑	DINO↑	CLIP-I↑	CLIP-T↑
4	2.19	0.692	0.820	0.292
8	**2.42**	**0.742**	**0.834**	**0.295**
12	2.09	0.715	0.826	0.283

Table 6: Ablation on modules in DRA-Ctrl.

	VL↑	DINO↑	CLIP-I↑	CLIP-T↑
Oracle	–	0.774	0.885	–
w/o loss reweighting	2.32	0.744	0.839	0.292
w/o FSPE	2.28	0.777	0.853	0.287
w/o mixup strategy	2.38	**0.900**	**0.918**	0.271
w/o attention masking	2.41	0.777	0.856	0.292
full version	**2.42**	0.742	0.834	**0.295**

Table 7: Generation efficiency analysis.

	latent frames	VL↑	DINO↑	CLIP-I↑	CLIP-T↑	Time/s↓
Oracle	–	–	0.774	0.885	–	–
I2V baseline	2	2.34	**0.803**	**0.874**	0.291	**10.8**
DRA-Ctrl	4	**2.44**	0.715	0.821	**0.298**	24.0
I2V	37	1.09	0.698	0.810	0.257	251

Comparison between T2V/I2V baselines. Quantitative results 3 on DreamBench align with Figure 3 and Section 3.2. The T2V baseline, whose base model is unable to accept images as control signals, achieves a high CLIP-T score but suffers from low DINO and CLIP-I scores. The I2V baseline produces condition image-like outputs, with the DINO score even surpassing the result measured on real images, but suffers from low prompt adherence. Under identical experimental configurations, DRA-Ctrl achieves a balanced performance, with DINO, CLIP-I and CLIP-T positioned between the two baselines and the highest VL Score, exhibiting superior performance.

Different mixup-based shot transition types. In addition to the fade-in-fade-out approach for constructing transition frames, we also experimented with slide-away transitions, with examples illustrated in Figure 6. Quantitative results in Table 4 demonstrate that the fade-in-fade-out mixup strategy outperforms slide-away across all three metrics. This observation aligns with our findings that video models tend to exhibit stronger priors for fade-in-fade-out shot transitions, while showing weaker priors for more complex transition types.

Number of transition frames. We investigate the impact of varying numbers of transition frames on experimental results, as shown in Table 5. Both insufficient and excessive transition frames harm performance. This phenomenon may stem from two factors: too few frames create excessively large inter-frame variations that increase learning difficulty, while too many frames introduce unnecessary computational overhead and slower convergence under the same training budget.

Module ablation. We conduct ablation studies on our proposed modules, including loss reweighting, FSPE, mixup strategy, and attention masking, with experimental results summarized in Table 6. Since our method employs an I2V model as the base architecture, all proposed modules aim to address its inherent limitations of excessive similarity to the condition image and poor prompt adherence. The results demonstrate that FSPE, mixup strategy, and attention masking significantly mitigate these issues, while loss reweighting primarily accelerates model convergence.

4.6 Generation Effiency Analysis

To analyze DRA-Ctrl's generation efficiency, we compare against the I2V baseline and the I2V model on DreamBench, assessing generation quality and efficiency. The I2V model generates videos from prompts and condition images, using the final frames as outputs. With $\delta = 12$ in FSPE (corresponding to 48 pixel-space frames), we set the I2V model to produce 145-frame videos. Table 7 results show our method achieves 90.4% faster generation than the I2V model with the highest VL score.

5 Conclusion

Leveraging the rich high-dimensional information priors inherent in video models, we propose to repurpose them for low-dimensional controllable image generation, demonstrating advantages akin to a "*Dimensionality-Reduction Attack*" effect compared to conventional image generation models. Specifically, to bridge the gap between video models' native capability for modeling continuous smooth transitions and the requirement for discrete abrupt changes in controllable image generation, we introduce a novel mixup-based transition strategy that constructs smooth transition between condition image and target image. Moreover, we redesign the attention masking mechanism that precisely aligns text prompts with image-level control signals. Our work establishes a new paradigm for activating high-dimensional video models to solve low-dimensional image generation tasks, while paves the way for future development of unified generative models across visual modalities.

Limitations. Our method employs a video model not optimized for image generation, resulting in slightly inferior performance on image quality metrics (FID, SSIM) compared to image-specific approaches. Besides, since HunyuanVideo-I2V primarily uses LLaVA [8] for prompt understanding, our CLIP-T scores are marginally lower than competing methods. Additionally, the requirement for transitional frames leads to reduced generation efficiency.

References

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

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL `https://arxiv.org/abs/2104.14294`.

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2021. URL `https://arxiv.org/abs/2012.00364`.

[5] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator, 2022. URL `https://arxiv.org/abs/2209.14491`.

[6] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024.

[7] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang Zhao. Unireal: Universal image generation and editing via learning real-world dynamics, 2024. URL `https://arxiv.org/abs/2412.07774`.

[8] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. `https://github.com/InternLM/xtuner`, 2023.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[10] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4744–4753, 2024.

[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, 2022. URL `https://arxiv.org/abs/2208.01618`.

[12] 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, 2024. URL https://arxiv.org/abs/2410.00086.

[13] 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, 2018. URL https://arxiv.org/abs/1706.08500.

[14] 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, 2021. URL https://arxiv.org/abs/2106.09685.

[15] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation, 2024. URL https://arxiv.org/abs/2311.17117.

[16] Miao Hua, Jiawei Liu, Fei Ding, Wei Liu, Jie Wu, and Qian He. Dreamtuner: Single image is enough for subject-driven generation, 2023. URL https://arxiv.org/abs/2312.13691.

[17] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers, 2024. URL https://arxiv.org/abs/2410.23775.

[18] Linyan Huang, Haonan Lin, Yanning Zhou, and Kaiwen Xiao. Flexip: Dynamic control of preservation and personality for customized image generation, 2025. URL https://arxiv.org/abs/2504.07405.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018. URL https://arxiv.org/abs/1611.07004.

[20] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. URL https://arxiv.org/abs/2212.04488.

[22] Nupur Kumari, Xi Yin, Jun-Yan Zhu, Ishan Misra, and Samaneh Azadi. Generating multi-image synthetic data for text-to-image customization, 2025. URL https://arxiv.org/abs/2502.01720.

[23] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[24] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, 2023. URL https://arxiv.org/abs/2305.14720.

[25] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning, 2025. URL https://arxiv.org/abs/2504.07960.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

[27] Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models, 2025. URL https://arxiv.org/abs/2503.10406.

[28] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[29] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling, 2025. URL https://arxiv.org/abs/2501.02487.

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. URL https://arxiv.org/abs/2108.01073.

[31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.08453.

[32] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models, 2024. URL https://arxiv.org/abs/2310.02992.

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

[34] 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.

[35] Senthil Purushwalkam, Akash Gokul, Shafiq Joty, and Nikhil Naik. Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models, 2024. URL https://arxiv.org/abs/2401.13974.

[36] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild, 2023. URL https://arxiv.org/abs/2305.11147.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[39] Noam Rotstein, Gal Yona, Daniel Silver, Roy Velich, David Bensaïd, and Ron Kimmel. Pathways on the image manifold: Image editing via video generation, 2025. URL https://arxiv.org/abs/2411.16819.

[40] 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, 2023. URL https://arxiv.org/abs/2208.12242.

[41] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

[42] Ke Sun, Jian Cao, Qi Wang, Linrui Tian, Xindi Zhang, Lian Zhuo, Bang Zhang, Liefeng Bo, Wenbo Zhou, Weiming Zhang, et al. Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person. *arXiv preprint arXiv:2407.16224*, 2024.

[43] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. URL https://arxiv.org/abs/2411.15098.

[44] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024.

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

[46] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance, 2025. URL https://arxiv.org/abs/2406.07209.

[47] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning, 2023. URL https://arxiv.org/abs/2212.02499.

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[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, 2023. URL https://arxiv.org/abs/2302.13848.

[50] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation, 2025. URL https://arxiv.org/abs/2504.02160.

[51] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. URL https://arxiv.org/abs/2409.11340.

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data, 2024.

[53] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

[54] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

[55] Xin Yu, Tianyu Wang, Soo Ye Kim, Paul Guerrero, Xi Chen, Qing Liu, Zhe Lin, and Xiaojuan Qi. Objectmover: Generative object movement with video prior, 2025. URL https://arxiv.org/abs/2503.08037.

[56] Yu Zeng, Vishal M. Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation, 2024. URL https://arxiv.org/abs/2407.06187.

[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

[59] Yabo Zhang, Xinpeng Zhou, Yihan Zeng, Hang Xu, Hui Li, and Wangmeng Zuo. Framepainter: Endowing interactive image editing with video diffusion priors, 2025. URL https://arxiv.org/abs/2501.08225.

[60] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, and Zhongliang Jing. Ssr-encoder: Encoding selective subject representation for subject-driven generation, 2024. URL https://arxiv.org/abs/2312.16272.

[61] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer, 2025. URL https://arxiv.org/abs/2503.07027.

[62] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2305.16322.

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. URL https://arxiv.org/abs/1703.10593.

[64] zk. text-to-image-2m (revision e64fca4), 2024. URL https://huggingface.co/datasets/jackyhate/text-to-image-2M.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: In the abstract and introduction (Section 1), our main claims reflect the paper's contribution and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in Section 5.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Our experimental details are provided in Section 4.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide open access to data and code.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide experimental details in Section 4.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: All comparison experiments are conducted under the same experimental settings.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Related experimental details are provided in Section 4.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our research fully complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss societal impacts of the work in the Appendix.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [Yes]

    Justification: We describe the safeguards in the Appendix.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All assets used in the paper are properly credited in compliance with academic standards

    Guidelines:

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

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We have released our assets.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: We use a Vision-Language model for evaluation, and details are provided in Section 4.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A More Experimental Details

In this section, we provide additional experimental details, including the configurations of LoRA and other hyperparameters. For different tasks, we employ distinct settings: Section A.1 describes the spatially-aligned image generation tasks, Section A.2 covers the subject-driven image generation task, and Section A.3 presents the experimental details for style transfer.

`DRA-Ctrl` employs LoRA [14] to fine-tune the base model with a rank of 16. Since our method needs to simultaneously process noiseless condition image token sequences and noisy generated image token sequences, we set the LoRA scale to 0 when handling the generated image token sequences to distinguish between them. Additionally, we set δ to 12 in the Frame Skip Position Embedding (FSPE). This configuration enables 4 frames in the latent space to effectively emulate 37 frames, corresponding to $1 + 36 \times 4 = 145$ frames in pixel space — approximately equivalent to a 5-second short video at 30 frames per second (fps), which sufficiently achieves the shot transition effect.

A.1 Spatially-aligned Image Generation

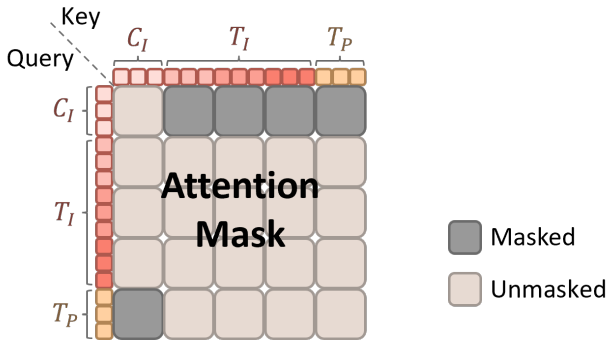


Figure 7: Attention masking strategy in spatially-aligned tasks.

In spatially-aligned image generation tasks, the condition image is directly extracted from the ground-truth image without a corresponding prompt. Therefore, we do not employ the condition image prompt C_P in our experiments, but we still utilize the attention masking strategy, with the corresponding attention mask illustrated in Figure 7. Besides, we train the model for 6,000 steps. In depth-to-image and depth prediction tasks, the depth image is extracted from the ground-truth image using Depth Anything [52]. For the depth prediction task, we prepend "[depth] " to the prompt to guide the model to generate depth maps rather than regular images. In the deblurring task, we apply Gaussian blur to the images with a randomly selected integer blur radius between 1 and 10 during training. For the in/out-painting task, we randomly select a rectangular region in the image during training, then mask either the selected region (with 0.5 probability) or the area outside it (with 0.5 probability) to create the condition image. In the super-resolution task, the condition image is obtained by downsampling the original image by a factor of 4.

A.2 Subject-driven Image Generation

For the subject-driven image generation task, we train the model for 9,000 steps. During inference, while employing attention masking, the simultaneous presence of both target image prompts T_P and condition image prompts C_P may still cause information blending. To address this, we strengthen the interaction between target image tokens T_I and T_P while suppressing C_P's influence on the generated output. Specifically, within the $(T_I \times T_P)$ attention mask region, we augment the attention weights by adding $0.6 \times \mu$ (where μ denotes the mean absolute value of the original weights). The modified attention computation for this region is formulated as:

$$\text{Attention}(Z) = \text{softmax}\left(\frac{Q_Z K_Z^\top}{\sqrt{d}} + 0.6 \times \text{mean}\left(\left|\frac{Q_Z K_Z^\top}{\sqrt{d}}\right|\right)\right) V_Z. \tag{4}$$

A.3 Style Transfer

Figure 8: Bitmoji-style example images in our dataset.

```
[USER PROMPT]:
将上传的图像分别转换为 bitmoji 风格，尺寸大小为{}:{}，输出清晰的图像。
```

Figure 9: The prompt format used for generating Bitmoji-style images with GPT-4o.

We collected 100 diverse images containing subjects such as humans, animals and buildings from the web. Using carefully designed prompts, we guided ChatGPT-4o to generate corresponding Bitmoji-style images, which formed our training set. The subject-driven image generation model is fine-tuned for 2,600 steps with a batch size of 8 on an NVIDIA H800 GPU to obtain the final model. Example images from our dataset are shown in Figure 8, and the prompt format we employed is shown in Figure 9, where the image dimensions are determined by their original resolutions.

B More Details about the VL Score

Current evaluation metrics for subject-driven image generation primarily employ DINO and CLIP-I to assess subject consistency, and CLIP-T for prompt adherence. However, two critical limitations exist: first, there lacks a comprehensive metric to directly evaluate subject-driven generation quality; second, these existing metrics exhibit notable shortcomings — both DINO and CLIP-I are significantly influenced by background interference, while CLIP-T struggles with fine-grained semantic alignment.

To address these issues, we propose leveraging an advanced Vision-Language (VL) model, such as QWen2.5-VL [1], as an evaluator to produce a holistic metric. Our approach consists of three steps: First, we provide the VL model with a prompt instructing it to score (prompt, reference image, generated image) triplets based on multiple fine-grained criteria for both subject consistency and prompt adherence. Next, we have the model summarize its task to confirm proper understanding. Finally, we input each triplet and collect the model's scores. Since both metrics are discrete scores ranging from 0 to 4, we average them to derive a comprehensive metric termed the VL Score. An example input-output demonstration of the VL model is shown in Figure 10.

C More Visualization

This section presents additional qualitative experimental results across all tasks, including transition frames generated by our model. The spatially-aligned image generation results are detailed in Section C.1, while the subject-driven image generation outcomes are presented in Section C.2, and the style transfer performance is analyzed in Section C.3. Unless otherwise specified, all image generation in this paper uses 50 sampling steps by default, including both qualitative results and quantitative evaluations, and generated images maintain a consistent resolution of 512×512 pixels.

```
[USER PROMPT]:
[Task Instruction] As a professional evaluator, your role is to assess the semantic consistency between a
↪  reference image and an generated image based on a given text prompt. Focus on two aspects:
1. Subject Consistency: Compare key visual attributes (Shape: main body outline, structure, and
↪  proportions. Color: saturation, hue, brightness, and distribution. Texture: realism, clarity, and
↪  aesthetic appeal. Facial: facial features, only apply when subject is a person or animal) between the
↪  reference and generated image.
2. Prompt Adherence: Verify if the generated image accurately reflects all elements described in the text
↪  prompt (Relevance: whether core elements in text are shown. Accuracy: objects, settings, actions, or
↪  characteristics Completeness: whether the generated image omits significant details. Context:
↪  portrayal of the appropriate environment, interactions, and background elements).
You need to provide your scores Subject Consistency and Prompt Adherence based on the following criteria.
[Subject Consistency Scoring Criteria] Very Poor (0): No resemblance. The generated image's subject has
↪  no relation to the reference.
Poor (1): Minimal resemblance. The subject falls within the same broad category but differs
↪  significantly.
Fair (2): Moderate resemblance. The subject shows likeness to the reference with notable variances.
Good (3): Strong resemblance. The subject closely matches the reference with only minor discrepancies.
Excellent (4): Near-identical. The subject of the generated image is virtually indistinguishable from the
↪  reference.
[Prompt Adherence Scoring Criteria] Very Poor (0): No resemblance. The generated image's subject has no
↪  relation to the reference.
Poor (1): Minimal resemblance. The subject falls within the same broad category but differs
↪  significantly.
Fair (2): Moderate resemblance. The subject shows likeness to the reference with notable variances.
Good (3): Strong resemblance. The subject closely matches the reference with only minor discrepancies.
Excellent (4): Near-identical. The subject of the generated image is virtually indistinguishable from the
↪  reference.
[Input Format] Every time you will receive two images and a prompt, the first image is a reference image,
↪  and the second image is the generated image, and the text is the prompt describing the generated
↪  image.
[Output Format] return the scores (0-4) in JSON format, strictly following this example:
↪  {\"subject_consistency\": 3, \"prompt_adherence\": 4}.

[VL MODEL OUTPUT]:
Sure, I understand the task. Please provide me with the reference image, the generated image, and the
↪  text prompt so that I can evaluate them according to the provided criteria.

[USER PROMPT]:
Did you understand the task above? Please summarize the tasks you need to do and show how you will
↪  execute the detailed plan for the task.

[VL MODEL OUTPUT]:
Yes, I understood the task clearly. too long, omitted

[USER PROMPT]:
{the first image} {the second image} {prompt}

[VL MODEL OUTPUT]:
```json\n{"subject_consistency": 3, "prompt_adherence": 2}\n```

More...
```

Figure 10: An example of VL Score evaluation process.

C.1 Spatially-aligned Image Generation Results

Our method performs image-to-video generation conditioned on input images, where the state of these condition images significantly impacts the output quality. We found that directly using canny edges, depth maps with black representing maximum depth, or black masks in in/out-painting tasks often resulted in unnaturally dark generated images. To address this, we implemented a color normalization scheme that remaps the darkest values $(0, 0, 0)$ to medium-gray $(128, 128, 128)$ while linearly scaling all other color values proportionally, preventing extreme darkening.
```

C.1.1 Canny-to-image

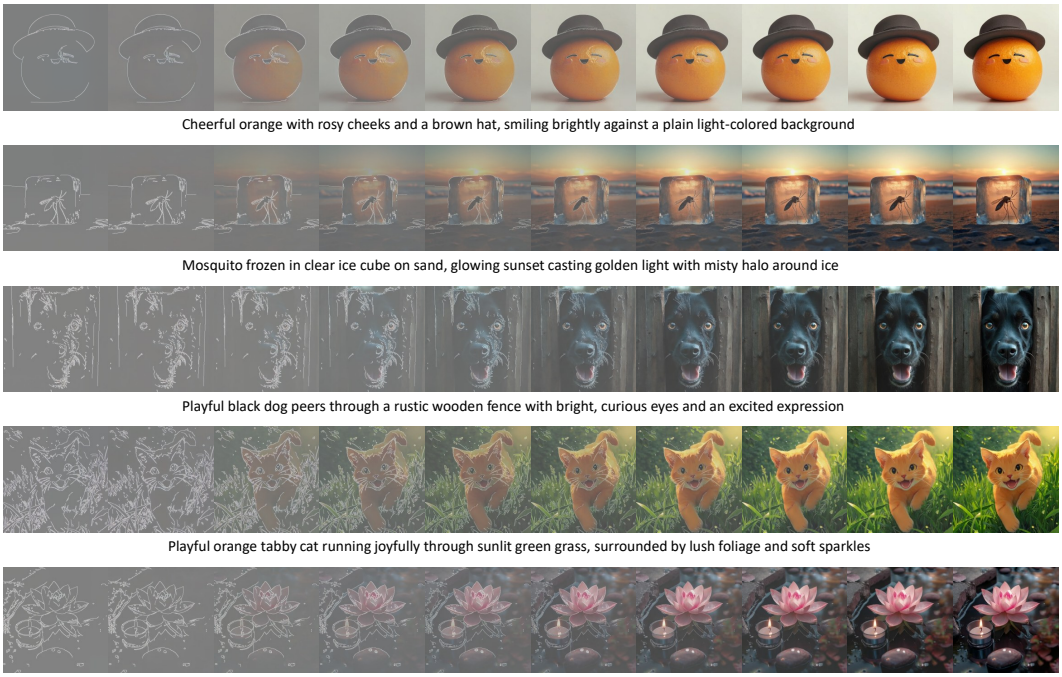


Figure 11: More canny-to-image generation results.

C.1.2 Colorization

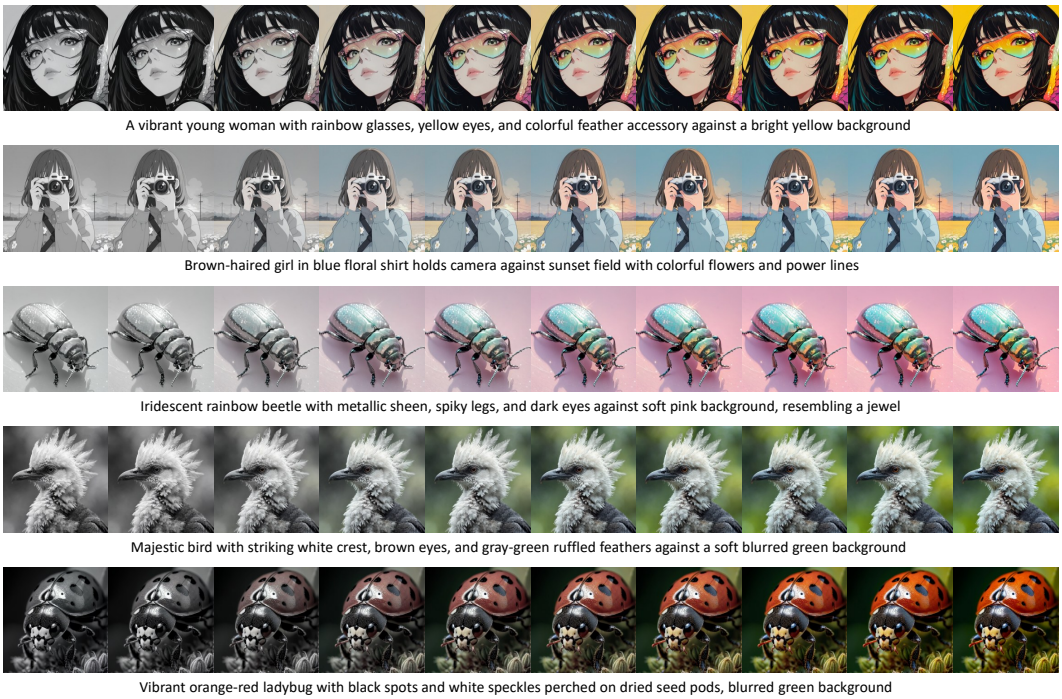


Figure 12: More colorization generation results.

C.1.3 Deblurring

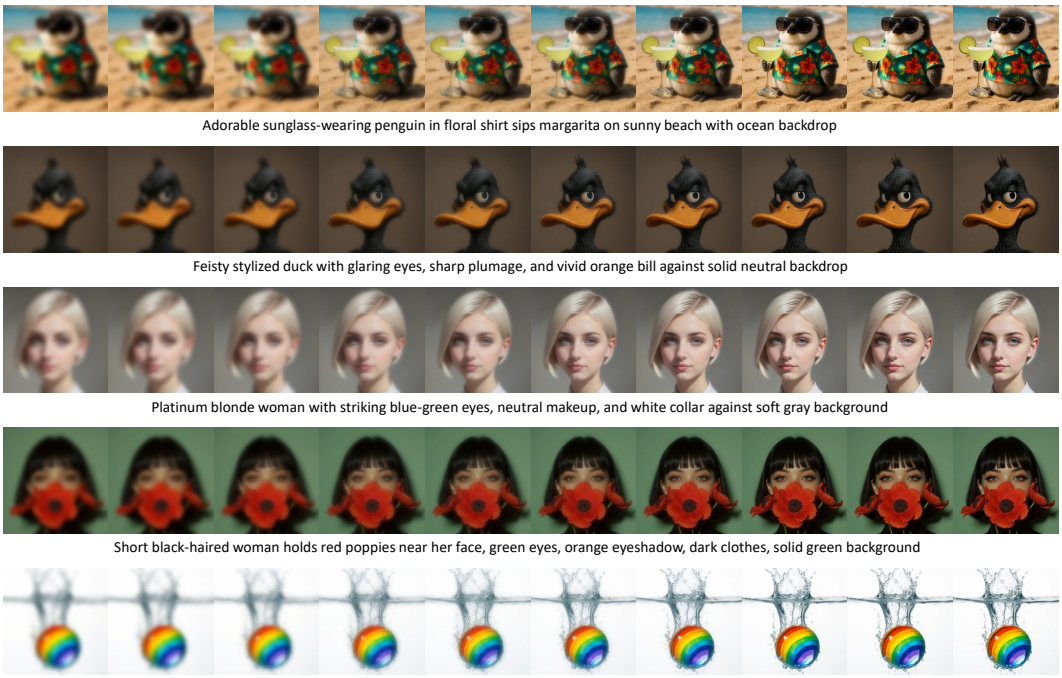


Figure 13: More deblurring generation results.

C.1.4 Depth-to-image

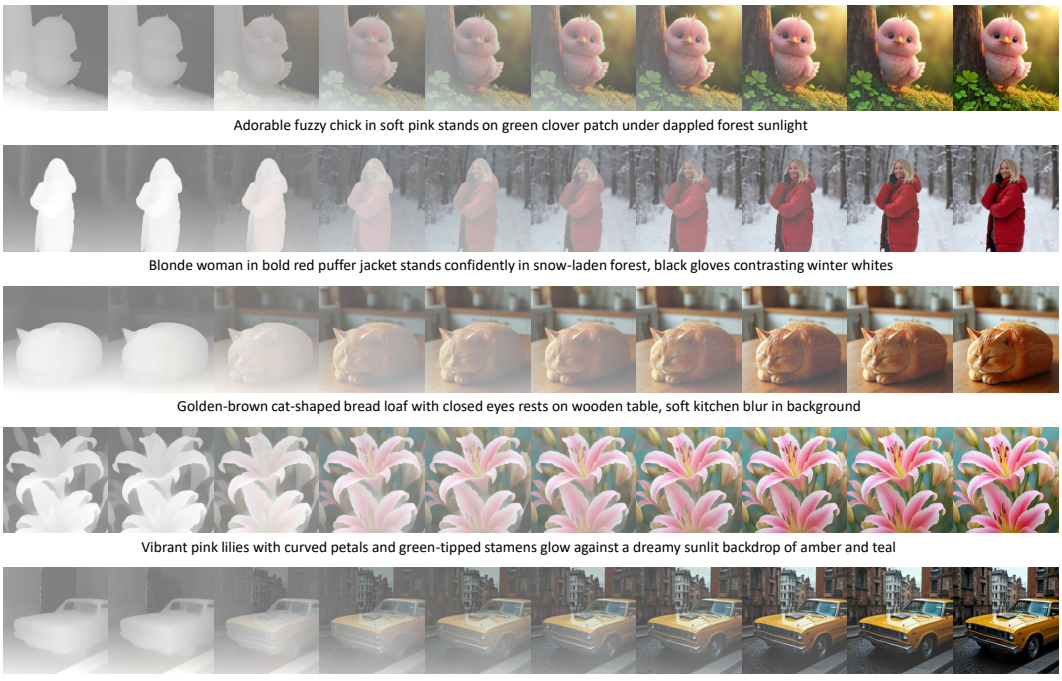


Figure 14: More depth-to-image generation results.

C.1.5 Depth Prediction

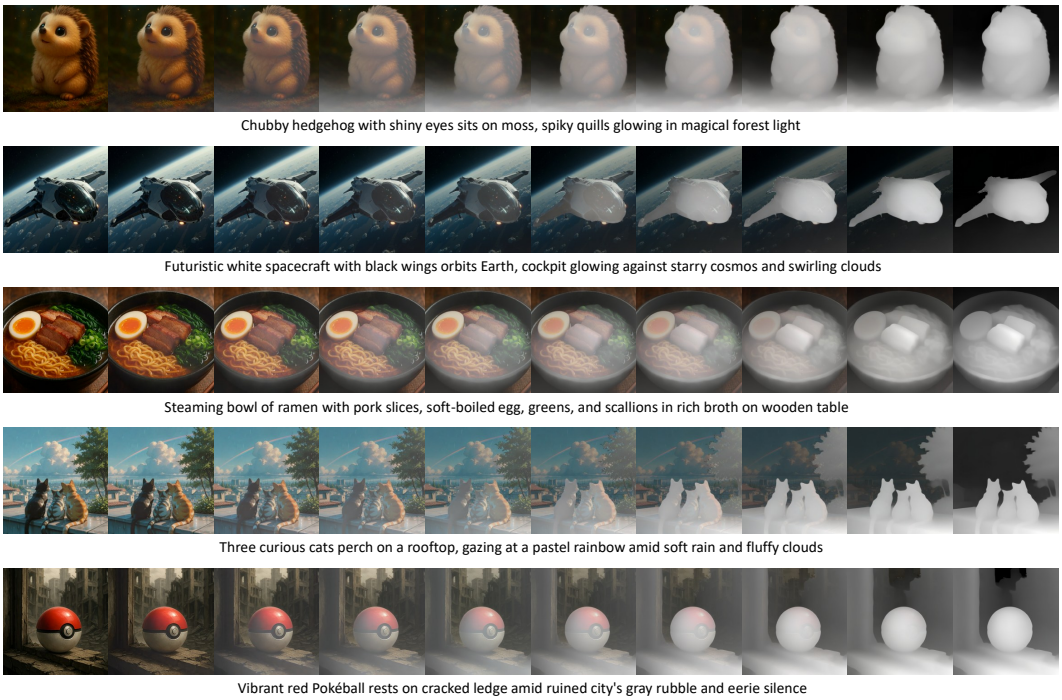


Figure 15: More image-to-depth generation results.

C.1.6 In/out-painting

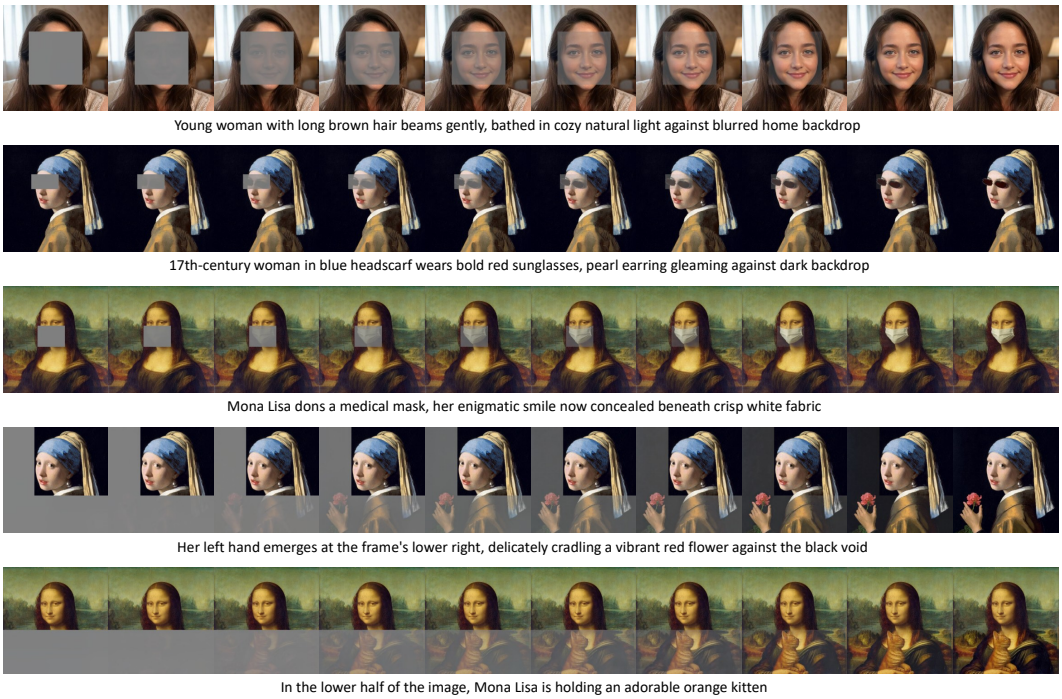


Figure 16: More in/out-painting generation results.

C.1.7 Super-resolution

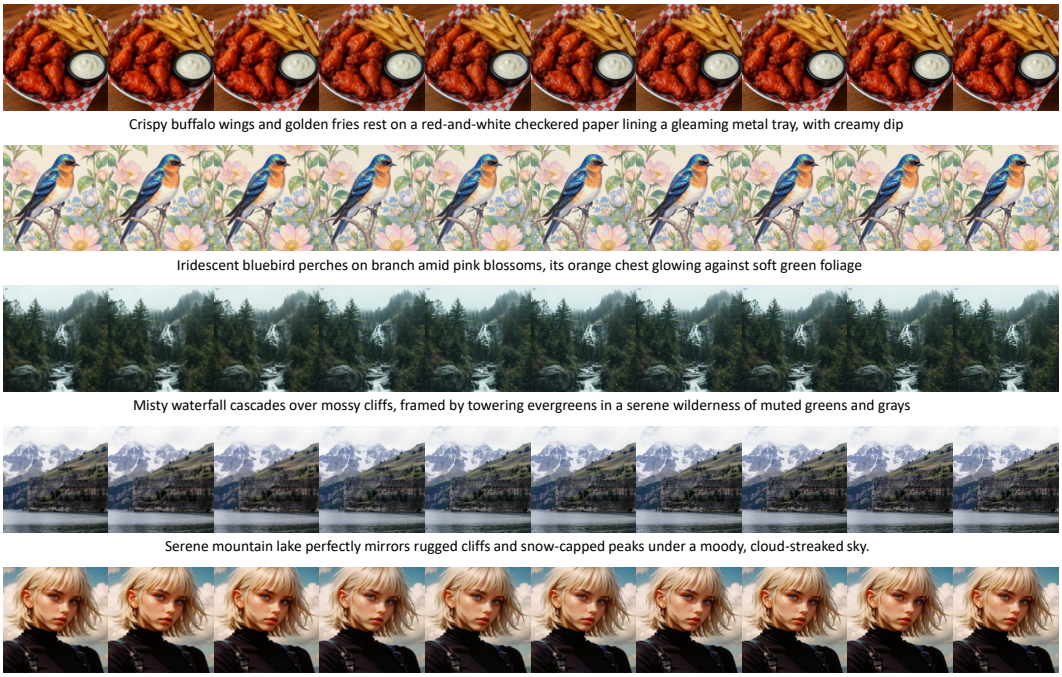


Figure 17: More super-resolution generation results.

C.2 Subject-driven Image Generation Results

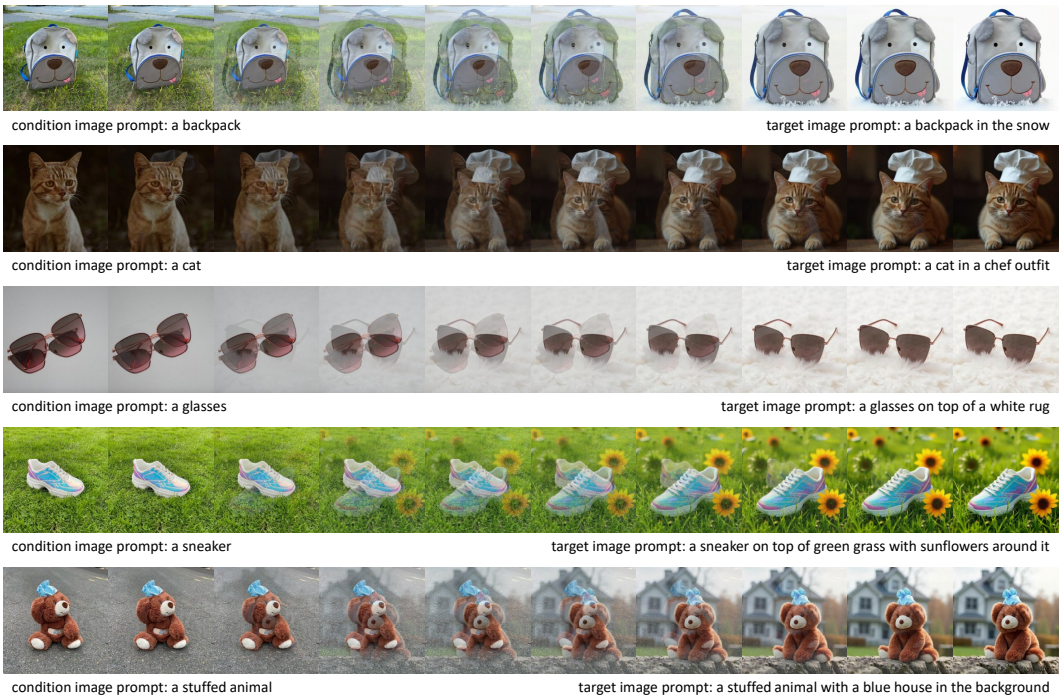


Figure 18: More subject-driven generation results on DreamBench.

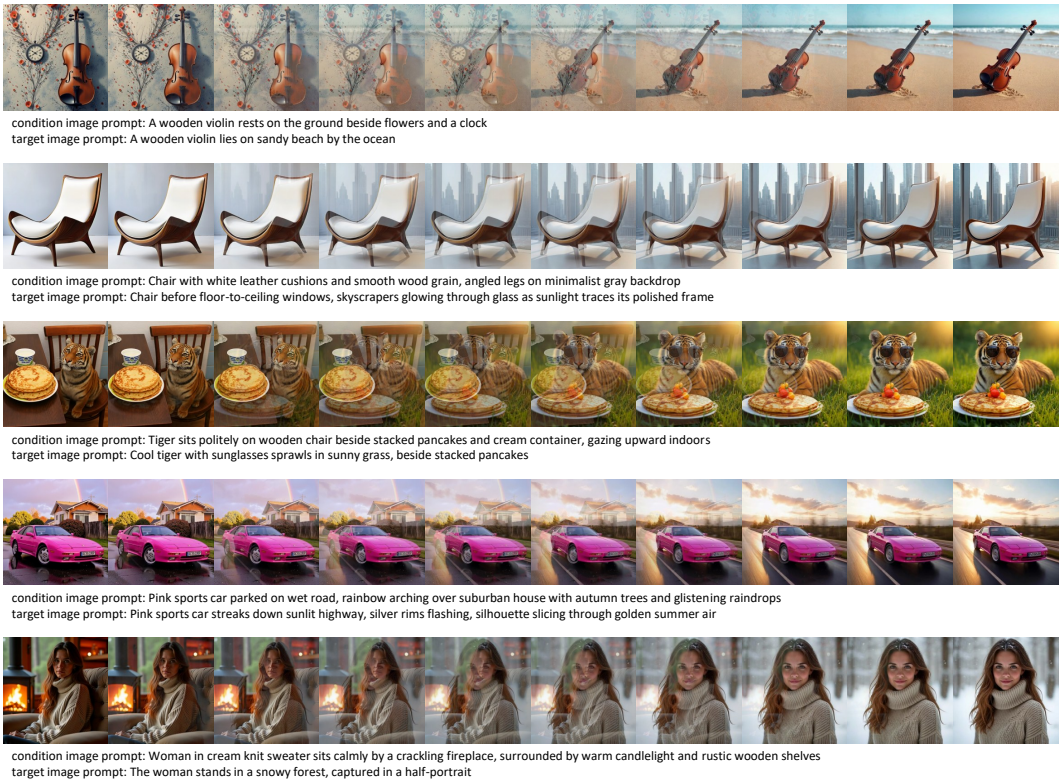


Figure 19: More subject-driven generation results.

Interestingly, we discover that during subject-driven image generation, `DRA-Ctrl` can occasionally control two subjects in the condition image simultaneously. As shown in the third row of Figure 19, our method successfully makes the tiger wear sunglasses while placing the stacked pancakes on the grass.

C.3 Style Transfer

Figure 20: More style transfer generation results.

D Failure Cases

While `DRA-Ctrl` successfully achieves controllable image generation in most cases, it may occasionally fail in the image-to-depth task, primarily manifesting as the presence of colored regions in the generated depth images. We attribute this limitation to the inherent nature of video models, which predominantly generate color data. A failure case is presented in Figure 21.

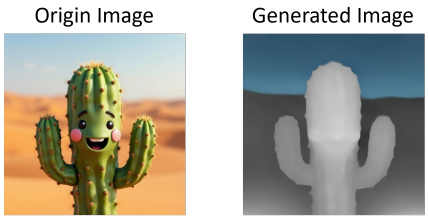


Figure 21: A failure case of `DRA-Ctrl`.

E Societal Impact

Our work advances controllable image generation with significant societal implications, offering both opportunities for innovation and risks requiring proactive mitigation. Below, we outline the potential positive and negative impacts, alongside measures to address the latter.

On the positive side, our high-quality, controllable generation method empowers creative and practical applications. Artists and designers can leverage it to produce imaginative content efficiently, while

educators benefit from dynamically generated visual aids for teaching. The fine-grained control also enables ethical uses in journalism and advertising, enhancing productivity and accessibility across domains.

However, negative impacts must be acknowledged. Malicious actors could exploit the technology to create convincing fake images for disinformation, fraud, or impersonation; to mitigate this, we adopt a gated release of models to restrict access. Bias in training data might lead to stereotypical or discriminatory outputs, disproportionately harming marginalized groups — addressed through rigorous bias testing during development. Further, misuse for non-consensual imagery (e.g., deepfakes) necessitates monitoring mechanisms and legal safeguards to protect privacy.

In summary, while our technology unlocks creative and educational potential, its risks-particularly around misinformation, bias, and privacy-demand deliberate countermeasures. By combining technical safeguards with policy-oriented solutions, we aim to foster responsible use and maximize societal benefit.

F Safeguards

To mitigate potential misuse risks associated with our controllable image generation technology, we will implement a gated release strategy when making the models publicly available. This will include: comprehensive usage guidelines explicitly prohibiting malicious applications such as disinformation campaigns and non-consensual imagery generation; an access control mechanism requiring users to agree to ethical use terms before obtaining the model. While we recognize no safeguards can eliminate all risks, these measures represent our proactive commitment to responsible AI development and deployment.