Tea-Adapter: Teacher Adapter for Efficient Conditional Generation

Yinhan Zhang[1,2*] Yue Ma[3*] Fangqiu Yi[2] Chenyang Qi[3] Chi Zhang[2] Zeyu Wang[1,3†]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]The Hong Kong University of Science and Technology

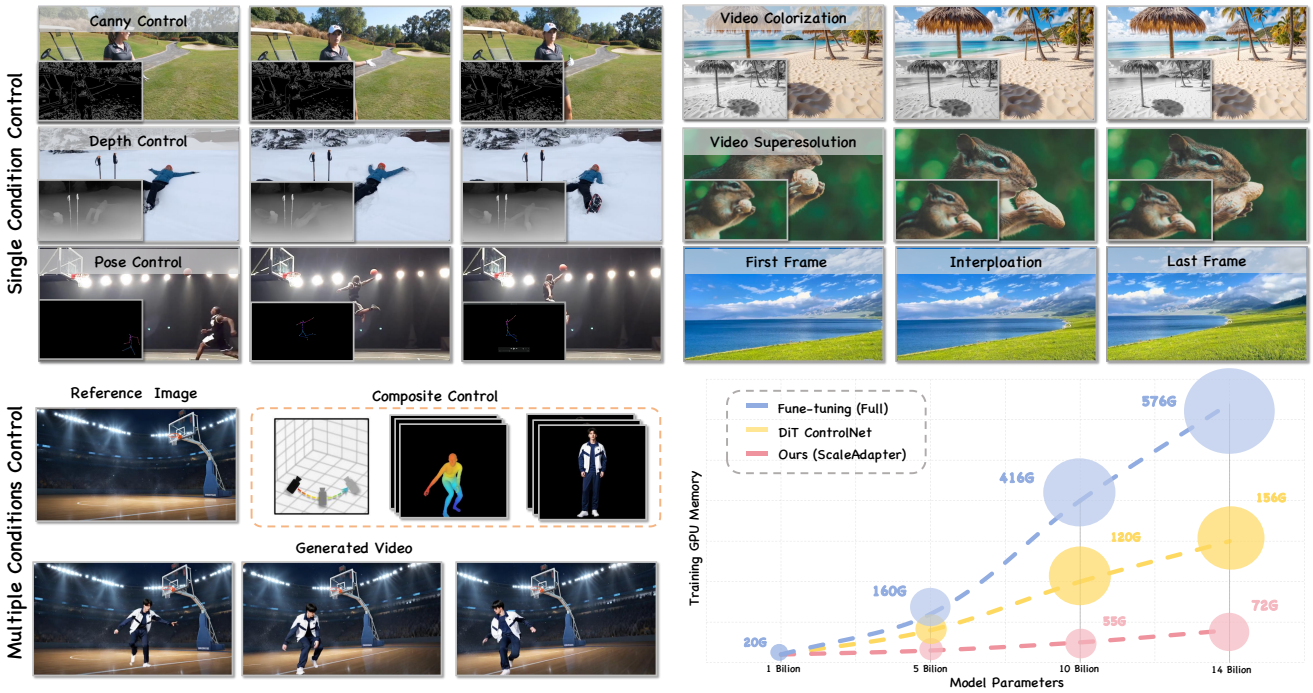

Figure 1. **Visual results of Tea-Adapter.** We propose the Tea-Adapter, a plug-and-play adapter that enables efficient training and flexible extension to diverse conditions, including both single conditions (e.g., canny, pose, depth) and multiple condition compositions (e.g., camera trajectory, image background, human motion) with minimal GPU consumption.

Abstract

We propose Tea-Adapter, a plug-and-play adapter designed to efficiently integrate conditional knowledge from a smaller teacher model into a larger student video diffusion model. Existing controllable video DiT methods face critical challenges: full fine-tuning of billion-parameter models is extremely expensive, while cascaded ControlNets introduce substantial parameter overhead and exhibit limited flexibility for novel multi-condition compositions. To overcome these issues, Tea-Adapter introduces a novel reverse distillation method that enables large video diffusion models to inherit precise control capabilities from smaller, efficiently tuned teacher diffusion models, eliminating the need for full fine-tuning. Moreover, recognizing the intrinsic relationships between different conditions, we replace the cascaded ControlNet design with a Mixture of Condition Experts (MCE) layer. This structure dynamically routes diverse conditional inputs within a unified architecture, supporting both single-condition control and multiple condition combinations without additional training cost. To achieve cross-scale knowledge transfer, we further de-

∗ Equal contribution.
† Corresponding author.
This work originated from an internship at TeleAI.

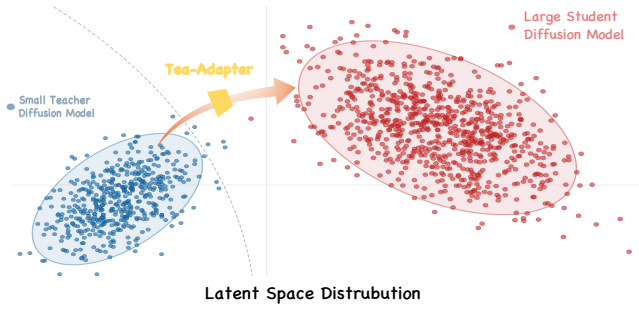

Figure 2. **The illustration of latent feature distribution transformation.** The design of our method is essentially about efficient feature transfer of control signals.

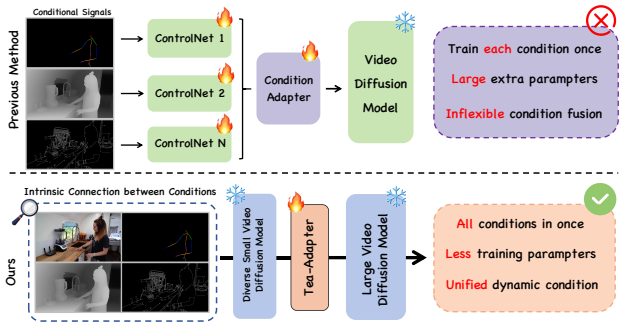

Figure 3. **Motivation of our method**. Compared to previous methods, our framework only requires training adapters in low-resource environments to support diverse conditions, eliminating redundant training efforts.

velop a Feature Propagation Module to ensure efficient and temporally consistent feature propagation across video frames. Experiments demonstrate that Tea-Adapter enables high-fidelity, multi-condition video synthesis, making advanced, controllable video generation feasible on low-resource hardware.

1. Introduction

Recent advances in Diffusion Transformers (DiTs) have revolutionized high-fidelity video synthesis driven by various conditions, enabling unprecedented visual quality [36]. However, generating spatiotemporally coherent content solely through text remains challenging due to the lack of guidance for fine-grained structural details (*e.g.*, object layouts, motion trajectories). To address this, control conditions, such as bounding boxes, segmentation maps, and depth maps, have been integrated into diffusion frameworks. Notably, ControlNet [59] and T2I-Adapter [34] have emerged as dominant solutions, extending Stable Diffusion [12, 37, 42] with lightweight adapters to support diverse input conditions, fostering broad adoption in controllable image generation. These methods enhance the conditional control capability of models by freezing the parameters of the main image generation network and introducing additional trainable parameters. When handling multiple conditional inputs, a common approach is to introduce additional ControlNets, each specialized for a specific condition [48]. However, this strategy leads to a linear increase in model parameters and requires repeated training processes for each new condition, resulting in significant computational overhead. Recent efforts Lin et al. [22] attempt to mitigate this by incorporating adapter modules and routers to combine multiple pretrained ControlNets for image diffusion models. Despite these improvements, such methods still rely heavily on existing pre-trained ControlNets, require substantial training resources, and exhibit limited extensibility to novel conditions.

Despite advancements in adaptive condition generation for video synthesis, current frameworks still face three critical challenges: (1) **Training Efficiency**: Fine-tuning ControlNets for DiT-based video diffusion models necessitates an enormous number of parameters. Incorporating a new conditional control typically requires about 0.5 billion parameters and significant computational resources (exceeding 48 GPU hours) for high-quality datasets. This imposes a substantial resource burden, compounded by the fact that state-of-the-art video models now have over 14 billion parameters. (2) **Inflexible Multi-Condition Fusion**: Prior works have extended conditional control from image generation to video generation tasks. While image-generation ControlNet architectures have been adapted for video tasks, they fail to address specific video control requirements, such as camera motion, background, or character features. Crucially, combining multiple conditions often relies on cascading specialized ControlNets. For a typical multiple conditions setup, it needs to train each separate ControlNet to learn each and equip it with a large base model (e.g., the Wan2.1 14B model), and the conditions cannot be dynamically integrated. (3) **Limited Condition Consistency**: Image-conditioned adapters often fail to maintain temporal and conditional coherence when applied to video generation, resulting in visible artifacts, including frame flickering and unstable content such as fluctuating characters or backgrounds. Although some extended approaches introduce temporal convolutions and linear projection layers into ControlNets, they still do not explicitly model the spatial correspondence and time-step alignment between conditioning features. This fundamental limitation necessitates large volumes of training data and extended training time, while still failing to ensure stable and controllable generation outcomes. At the same time, as shown in Figure 2, we noticed that small and large models within the same architecture family exhibit strong feature similarity (detail in the supplementary materials), allowing knowledge, particularly in

latent space, to be efficiently transferred from a fine-tuned small model to a large pretrained foundation model. Moreover, we note that low-resource fine-tuning can equip small models with richer and more diverse conditional control abilities than those achieved by single-condition ControlNet adaptations.

To overcome these limitations, we propose a unified framework centered on three innovations: First, we introduce Tea-Adapter, a novel and efficient method developed with minimal training overhead for conditional video synthesis. Instead of training a full video model, it enables synergistic collaboration between large foundation models and small video diffusion models, as shown in Figure 3. Unlike previous methods, we have found that small models and large models with the same architecture can achieve cross-scale knowledge bridging. Second, we design a *Mixture of Condition Experts* (MCE) that concurrently processes heterogeneous input conditions for video generation tasks in a single forward pass, and we find the intrinsic relationships between different visual conditions, thereby eliminating repeated training cycles. Unlike previous adapter-based methods, MCE employs dynamic routing to activate relevant experts for conditions, leveraging inter-condition synergies learned during joint training. As shown in Figure 6, this results in fewer parameters compared to Multi-ControlNet [48]. Third, we develop a Feature Propagation Module to ensure feature reversed propagation. Conditional features from the adapter are scaled and projected into each video DiT block, aligning the injected controls with the base model's priors. Our approach reduces condition-specific video training costs, supports dynamic composition of novel conditions, and cuts parameter overhead versus ControlNet and adapters, setting a new efficiency framework for controllable video generation. Our contributions are summarized as follows:

- We present Tea-Adapter, a plug-and-play adapter to transfer the controllable knowledge from small-parameter video models to large-parameter models efficiently.
- Technically, we first design a *Mixture of Condition Experts* (MCE) layer that covers various control signals with dynamic expert routing. It shows the ability to adapt to unseen conditions by learning the intrinsic relationships between different visual conditions.
- To achieve reversed condition distillation, we develop a *Feature Propagation Module* that efficiently ensures condition coherence during feature transfer in the denoising stage.

2. Related Work

Video Diffusion Model Generative modeling has propelled remarkable advancements in large-scale video models, with diffusion-based frameworks emerging as a prominent area of development [1, 6, 9, 11, 17, 24, 26, 29, 31, 33, 35, 57]. A large number of diffusion-based video generation approaches are built upon the Stable Diffusion [2, 30, 42], encompassing three fundamental components: an autoencoder that transforms raw videos into a compact latent space [50]; a text encoder tasked with extracting text embeddings [40]; and a neural network, optimized through diffusion processes, [13, 18, 39] that learns the distribution characteristics of these video latents. In terms of architectural design, the U-Net, originally devised for image generation tasks, has been adapted to video generation by integrating temporal dimensions. Notably, Diffusion Transformers (DiTs)[4, 10, 25, 36, 44, 55], which employ exclusively transformer blocks, have exhibited superior performance over U-Net architectures in the domain of visual generation.

Controllable Generation in Diffusion Models The remarkable success of diffusion models [3–5, 8, 14, 19–21, 23, 27, 28, 32, 45–47, 51–53, 58, 61, 62] has spurred substantial interest in controllable video generation. To address the need for fine-grained control over diffusion-based synthesis, researchers have explored a wide range of conditional inputs, including depth maps, Canny edges, reference images, and multimodal combinations. However, the computational cost of full-parameter fine-tuning for each new condition has driven the development of parameter-efficient adaptation methods. Notable approaches in this domain include ControlNet [59] and T2I-Adapter [34], which enable pretrained diffusion models to incorporate additional conditional signals through lightweight trainable branches. These methods effectively balance expressiveness and efficiency. Those methods [7, 16] are a conditional generation framework that constrains diffusion models to generate content aligned with structural control signals (e.g., sketches, depth maps) while preserving high generative quality.

UniControl [38] introduces a MoE-style Adapter and a Task-aware HyperNet to support diverse tasks within a single model. However, its task adaptation mechanism is designed for text instructions and does not explicitly model relationships between tasks and conditions. Multi-ControlNet [48], which enables composite control, suffers from isolated branches that limit composability. Uni-ControlNet [63] addresses this by grouping conditions into local and global controls, supporting composable control within a single model. Nevertheless, its inability to maintain consistency across frames hinders its applicability to video generation. Inspired by ControlNet, DiT-ControlNet [7] incorporates zero modules into the DiT architecture to learn new conditions without training the backbone model. While effective, this approach incurs significant training overhead. Ctrl-Adapter [22] injects latent feature maps into video generation models using image ControlNets and adapters inserted into each DiT block. However, it struggles to main-

tain temporal consistency across video frames. By contrast, Tea-Adapter takes only a single condition while still being capable of both multi-condition and zero-shot learning.

3. Method

3.1. Task Definition

Given a text description T, diverse visual conditions C, large text-to-video diffusion model F_l, conditional small video diffusion model F_s, the goal of Tea-Adapter S is to transfer various control signal guided generation ability in F_s to F_l without additional ControlNet training. A core requirement of S is that V_{gen} aligns with both text description T and diverse visual conditions C. Formally, this conditional video generation task is formulated as:

$$V_{\text{gen}} = F_l\left(T, S(F_s(C))\right). \tag{1}$$

Our designs are detailed in subsequent sections: Section 3.2 presents the overall architecture of $\mathcal{F}$, outlining the interaction mechanism between the Small Video Diffusion Model, Tea-Adapter, and Large Video Diffusion Model. Section 3.3 describes the details of adapter design, which enables efficient transfer of control conditions across different model scales.

3.2. Tea-Adapter Training Strategy

The framework in Figure 4 illustrates our methodology for enabling efficient transfer of scalable multi-condition control knowledge. Tea-Adapter propagates the knowledge from a small video diffusion model to a large one, keeping the parameters of both pretrained models frozen. Notably, the small video diffusion model, initialized from a pretrained text-to-video diffusion model, requires prior fine-tuning or LoRA training [15] to adapt to multiple conditions. Since the large video diffusion model has a different number of DiT blocks than the small model, we select the first, last, and several middle DiT blocks to transfer conditional latent features to the large model. This design allows us to train only the Tea-Adapter, resulting in significantly higher efficiency compared to fine-tuning the large model itself.

3.3. Adapter Architecture

Cross-Scale Knowledge Bridging. Inspired by ControlNet, conditional information is effectively injected into the target backbone using trainable copies of the diffusion model block and zero-initialized linear layers. As illustrated in Figure 4, our Tea-Adapter for the DiT-based video diffusion model employs a novel architecture consisting of three key components: attention modules, *Mixture of Condition Experts*(MCE), and a *Feature Propagation Module*. Leveraging the MCE layer to dynamically route condition tokens and adapt the timestep embedding t derived from the small video diffusion model within the Feature Propagation Module, our design ensures consistent conditional and temporal representation throughout the bridging stage of the diffusion process. At the same time, consider increasing the minimum number of additional parameters.

Mixture of Condition Experts. Observing that intrinsic connections exist among different conditioning signals, such as between canny edges and depth maps shown in Figure 3, we are inspired to develop a unified architecture that leverages these relationships for multi-condition generation. To avoid the inefficiency of retraining for each new condition while ensuring high scalability and zero-shot adaptation capability to unseen conditions, we introduce a *Mixture of Condition Experts* (MCE) layer. This module comprises a specialized set of experts within Tea-Adapter that work together to capture and integrate latent features from diverse conditional inputs, such as depth maps, Canny edges, and human poses. Within the MCE layer, different experts are designed to simultaneously learn from various conditional signals. Only a sparse subset of experts is activated during processing, enabling effective and efficient fusion under both single and multi-condition settings. This structure further allows the model to exhibit zero-shot generalization to new, unseen conditions during inference. Moreover, the MCE layer offers high extensibility. When introducing a new condition or task, new experts can be seamlessly added. These experts can be initialized using weights shared from existing experts, facilitating rapid convergence with minimal training data.

Our MCE layer consists of two types of parameterized experts: shared experts $\mathcal{E}_s$ and condition-specific experts $\mathcal{E}_c$. This design enables zero-shot generalization to unseen conditions by leveraging knowledge from related expert modules. Mathematically, given a set of K feature tokens $\{c_1, c_2, \ldots, c_K\}$, the MCE layer computes the conditional output h_t^{mce} in timestep t as:

$$h_t^{mce} = \sum_{k=1}^{K} g_k(c_k, t) \cdot \mathcal{E}_{c_k}(x_t^a, t), \tag{2}$$

where $g_k(c_k, t) = \text{Softmax}(\text{MLP}_g([c_k; t]))$ denotes the gating function that assigns weights to each expert $\mathcal{E}_{c_k}$ based on the input condition c_k and x_t^a is adapter attention module output. The shared expert $\mathcal{E}_s$ is integrated via:

$$\mathcal{E}_{c_k}(x_t^a, t) = \mathcal{E}_s(x_t^a, t) + \Delta\mathcal{E}_{c_k}(x_t^a, t), \tag{3}$$

where $\Delta\mathcal{E}_{c_k}$ represents the condition-specific adaptation parameters. During inference, only the relevant experts are activated via dynamic routing, reducing computational overhead. Empirically, this design achieves state-of-the-art multi-condition synthesis with fewer parameters compared to naive Multi-ControlNet baselines while maintaining condition and temporal consistency across frames.

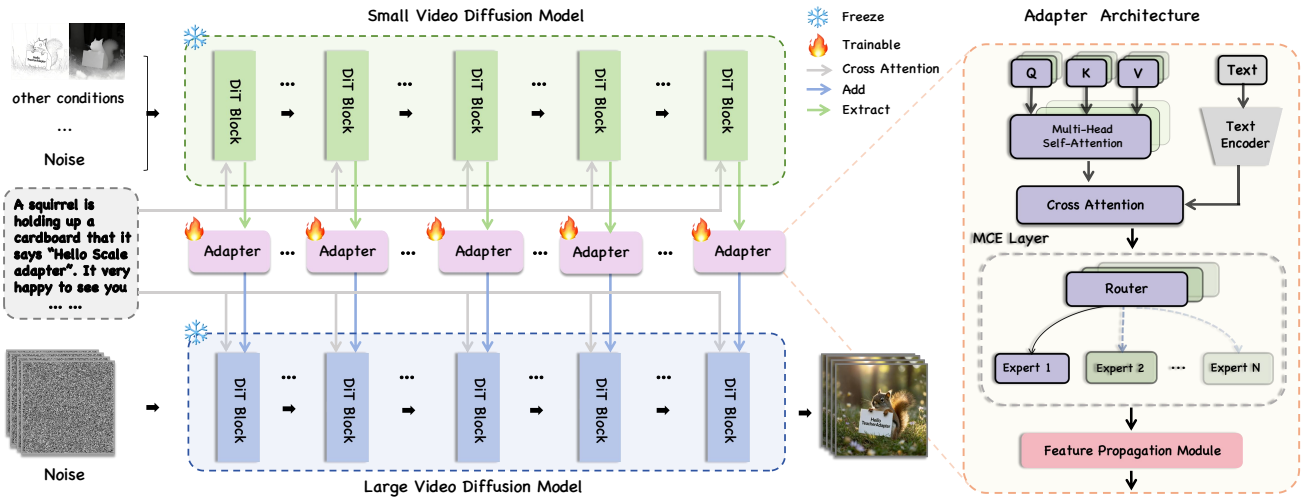

Figure 4. **Overview of Tea-Adapter**. **Left**: To drive a large text-to-video diffusion model with new conditions, we first feed the condition latents to a frozen small pretrained conditional diffusion model, whose features are first injected into Tea-Adapter and then mapped to the frozen large diffusion model. **Right**: For each adapter, we design an *Mixture of Condition Experts* (MCE) layer to learn multiple control signals and a Feature Propagation module to transfer knowledge efficiently.

Feature Propagation. To efficiently transfer condition information from the large model to the small model, we introduce a *Feature Propagation Module* that includes a learnable modulation factor, a time projection layer, and an Up-Projection layer. The Up-Projection layer leverages a linear layer to transfer condition information from the small video diffusion model to the large video diffusion model. Then, the learnable scaling modulation with a time projection layer dynamically adapts condition features into the large video diffusion model. Those designs enable Tea-Adapter to modulate latent feature contributions based on the denoising stage adaptively. Specifically, given the adapter's latent feature x_t^a at timestep t, we compute the cross-attention output y between x_t^a and the text embedding c_{txt}. The feature propagation process is formalized as follows:

$$\begin{aligned} \alpha_{\text{scale}} &= \text{Modulation} + \text{Time\_Proj}(t), \\ x_t^{a'} &= \text{Up\_Proj}(x_t^a) \cdot \alpha_{\text{scale}} + \text{Up\_Proj}(h_t^{mce}), \end{aligned} \tag{4}$$

Where α_{scale} denotes a learnable scale modulation factor that the latent feature transfers to the target video backbone. The overall feature propagation function is defined as

$$x_t^{a'} = u(x_t^a, c_{txt}, t; \theta), \tag{5}$$

where $x_t^{a'}$ represents the adapter output and θ encompasses the adapter's trainable parameters. The scale features are then integrated into the large model's latent space via:

$$x_t = x_t + x_t^{a'}, \tag{6}$$

Where x_t is the latent large video diffusion model during the denoise stage. This additive integration ensures that the large model's prior knowledge is augmented with condition-specific information while preserving its structural integrity. Through this module, Tea-Adapter effectively bridges the gap between small, specialized models and large foundation models, enabling efficient knowledge transfer across scales. Empirically, it reduces trainable parameters by 70% (without MCE layer) compared to DiT-ControlNet while maintaining comparable performance.

4. Experiments

4.1. Implementation Details

Tea-Adapter integrates Diffusion Transformer Blocks with Multiple Condition Experts (MCE) Layer [43]. We conduct experiments using two open-source text-to-video diffusion models as backbones: Wan2.1-1.3B and Wan2.1-14B [51], as well as CogVideoX-2B and CogVideoX-5B [14]. Training required approximately 2 days on 1×NVIDIA H100 80GB GPU. We sampled 15K videos from the Koala-36M dataset [54] and generated degraded versions by converting samples to grayscale and downscaling to low resolution. Before training, we extracted auxiliary conditioning signals (human pose, depth maps, and Canny edges) from all videos. For evaluation, we manually curated 100 high-quality videos spanning diverse content categories. For conditional generation tasks involving reference videos, we report LPIPS [60], SSIM [56], CLIP Score(semantic correspondence between generated and reference content), and FVD metrics [49].

Figure 5. **Qualitative comparisons with baselines.** "Ctrl" stands for "ControlNet" and "Apt" stands for "Adapter." We perform the visual comparison with five baselines using the same conditions, while the image-based method shows poor performance in cross-frame consistency, and our method obtains better performance in the adapter-based methods.

Table 1. Comparison of state-of-the-art baselines. The best result in each column is **bolded**, and the second best is underscored.

Model	Canny Edge				Depth Map				Pose				Temporal Consistency ↑
	FVD↓	CLIP↑	LPIPS↓	SSIM↑	FVD↓	CLIP↑	LPIPS↓	SSIM↑	FVD↓	CLIP↑	LPIPS↓	SSIM↑	
X-Adapter [41]	-	0.545	0.736	0.209	-	0.517	0.759	0.127	-	-	-	-	0.754
Uni-ControlNet [64]	-	0.642	0.575	0.322	-	0.531	0.778	0.214	-	0.509	0.823	0.188	0.763
UniControl [38]	-	0.584	0.773	0.268	-	0.572	0.791	0.178	-	0.541	0.741	0.207	0.876
Ctrl-Adapter [22]	427.060	0.757	0.358	0.619	448.291	0.785	0.352	0.616	487.429	0.712	0.672	0.304	0.981
DiT-ControlNet [7]	425.249	0.781	0.551	0.369	540.573	0.729	0.686	0.304	537.123	0.645	0.777	0.295	0.978
Wan2.1-14B(fine-tuned)	**229.186**	**0.919**	**0.187**	**0.675**	**254.238**	0.912	**0.193**	**0.664**	**200.911**	**0.926**	**0.161**	**0.691**	0.979
Ours	289.565	0.918	0.255	0.585	292.341	**0.913**	0.251	0.591	300.582	0.903	0.273	0.573	**0.984**

4.2. Qualitative Results

We visually compare the performance of our method against baseline models across three key conditions (Canny edges, Depth maps, and Openpose skeletons) in Figure 5. Our approach consistently outperforms alternatives in both visual quality and alignment with input conditions and text prompts, as validated by the qualitative examples. Under pose control, our method achieves significantly tighter spatial alignment with input skeletons compared to baselines. For the prompt "a woman walking along the shoreline" on Figure 5, our results accurately adhere to pose constraints while maintaining natural motion. In contrast, UniControl and Uni-ControlNet misinterpret skeletal configurations. For instance, the blue-dressed woman in their outputs exhibits inconsistencies in appearance and motion coherence, with subtle frame-to-frame discrepancies undermining temporal consistency. Our model, by contrast, preserves precise pose adherence while ensuring smooth, natural movements. For depth control generation, our framework demonstrates a superior understanding of 3D geometry, producing outputs with geometrically plausible structures from depth maps and surface normals. Ctrl-Adapter, by comparison, exhibits noticeable geometric inconsistencies, such as distorted character proportions and implausible spatial relationships between objects. The X-Adapter captures basic human semantics but suffers from significant inter-frame variations in character appearance, lacking the temporal consistency necessary for video generation. Our method, however, maintains both geometric fidelity and cross-frame coherence. For edge-guided generation, our model outperforms ControlNet-based methods in edge preservation and structural consistency. As shown in Figure 5, this advantage is particularly pronounced in motion details, for example, the leg movements of the yellow cheetah in the examples, where competing methods exhibit noticeable blurring or edge misalignment. Our results retain sharp, faithful alignment with input edges while pre-

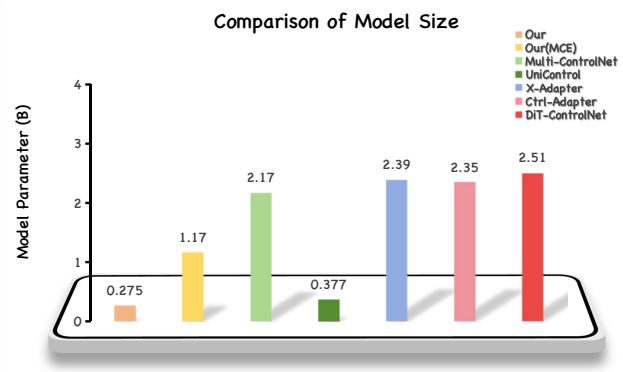

Figure 6. **Comparison of trainable model parameters in the diffusion model**. Our methodology requires the fewest trainable parameters for multiple control signals when the MCE layer is removed, compared with other methods.

serving the fluidity of dynamic motions. These qualitative findings reinforce that the integrated design of our method, combining the condition fusion of the MCE layer and efficient feature propagation, better balances condition adherence, visual quality, and temporal consistency between various control signals.

4.3. Quantitative Results

We conducted comprehensive comparisons against state-of-the-art ControlNet-based and Adapter-based methods. As shown in Table 1, Tea-Adapter, deployed on the 14B-T2V base model, outperforms existing strong video control methods across both depth map and Canny edge conditions, achieving competitive performance in visual quality and similarity to the reference video. For Adapter-based methods, our model outperforms X-Adapter and Ctrl-Adapter across most metrics, despite notable differences in training resources: while these baselines utilize datasets of over 100K videos or images and more GPUs, our method is trained on a more compact 10K video dataset. Notably, our base model lacks prior conditioning capabilities, highlighting the efficiency of our adapter design in injecting control capabilities into pre-trained text-to-video models. For ControlNet-based methods, we further compared against UniControl and Uni-ControlNet. We keep their method setting using the image diffusion model to generate frames. Tea-Adapter achieves the competitive FVD, LPIPS, and CLIP scores. In contrast, existing ControlNet-based approaches require distinct ControlNets for different conditions, allowing specialized training for each task, yet our method still achieves overall competitive performance. Additionally, we provided a comparative experiment of the same model architecture for small models, large models, and our method in different tasks in the supplementary materials. These results validate that Tea-Adapter's architecture delivers great reversed distillation ability for conditional video generation, balancing efficiency and performance across diverse metrics.

Table 2. Ablation study of key components.

Configuration	FVD ↓	LPIPS ↓	SSIM ↑	CLIP ↑
Full Model	**292.341**	**0.251**	**0.591**	**0.913**
w/o MCE	303.202	0.268	0.573	0.904
w/o Half of adapters	398.013	0.355	0.567	0.875

Figure 7. **Ablation results**. We present results by removing the MCE layer and changing the number of adapters. Without the MCE layer and a half number of adapters, it exhibits different levels of degradation in motion coherence and quality.

4.4. Ablation Study

We systematically evaluate the core components of Tea-Adapter through controlled ablation experiments, utilizing five metrics: FVD, LPIPS, SSIM, and CLIP Score. Quantitative comparisons in Table 2 reveal two key insights:

MCE Layer Efficacy The full model with the *Mixture of Condition Experts* (MCE) layer achieves an FVD improvement across metrics compared to MCE-ablated variants, underscoring the critical role of dynamic condition feature fusion. This improvement stems from mixed training on multiple conditions, which enables the model to learn intrinsic relationships between different conditions. While the MCE-ablated variant retains basic conditioning control, it exhibits measurable degradation in motion coherence. As shown in Figure 7, the MCE-equipped model produces clearer and smoother details in video characters (*e.g.*, hands and legs). Moreover, in scenarios with multiple characters (left panel of Figure 7), the MCE layer better controls individual character motions and their interactions with objects.

Adapter Scaling Efficiency As depicted in Figure 6, reducing the number of adapters from 12 to 7 maintains robust conditioning fidelity; most metrics show no significant degradation, while reducing adapter parameters by nearly half. This indicates that the combined design of the Feature Propagation Module and MCE layer effectively mitigates performance drops even with fewer adapters, making

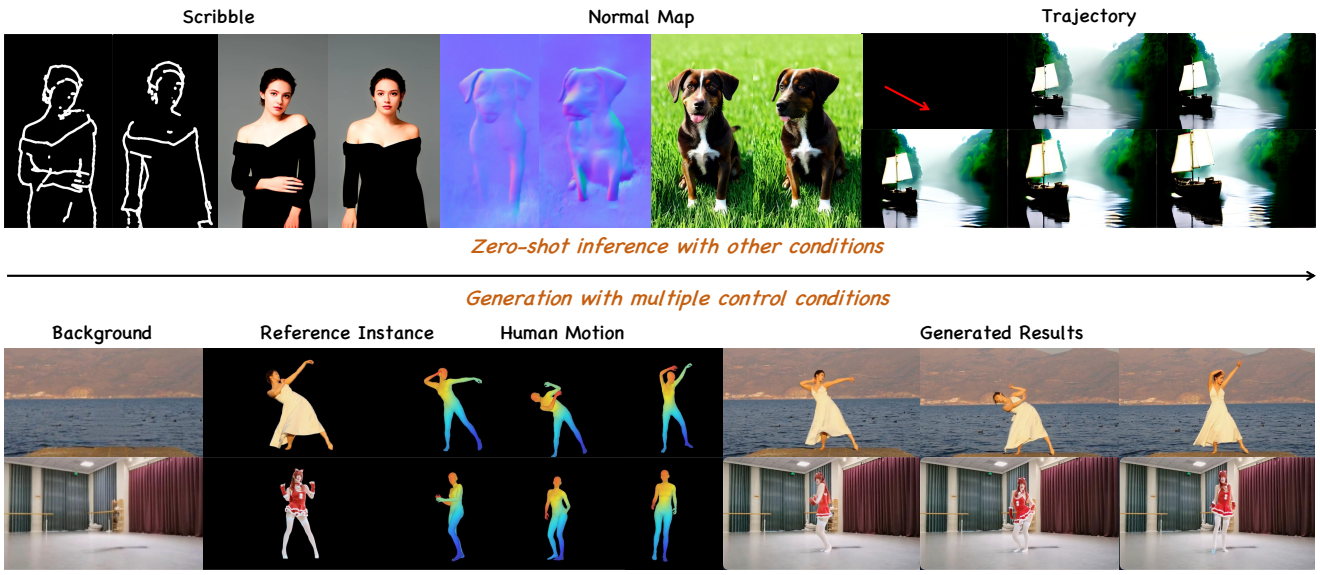

Figure 8. **Zero-shot and Multi-Condition Generation.** (Top) **Zero-Shot Generation:** our method, trained on one condition, generalizes to unseen conditions and produces high-quality, controllable videos. (Bottom) **Multi-Condition Generation:** Given multiple control signals (e.g., background, reference instance, and human motion), our approach can effectively transfer different conditions.

the model less sensitive to adapter count. However, further reducing the number of adapters leads to noticeable declines in video quality, particularly in condition consistency. These results demonstrate that the synergistic integration of the MCE layer and adaptive adapter scaling achieves state-of-the-art efficiency performance tradeoffs. This design enables precise conditional control with only single-pass inference, balancing model lightness and control capability. The additional zero-shot quantitative experiment is in the supplementary materials.

5. Applications and Discussion

Zero-Shot Generalization. Tea-Adapter acts as a parameter-efficient knowledge bridge, mapping diverse conditional signals into the backbone's unified representation space without full model fine-tuning. To evaluate generalization, we deployed it to large video diffusion models and tested on unseen conditions (scribble, normal map, trajectory) not included in training. Quantitative results in supplementary materials and qualitative examples in Figure 8 confirm its effectiveness: the model follows scribble contours, preserves normal map depth consistency, and generates trajectory-aligned smooth motion. This zero-shot capability eliminates the need for re-training or specialized annotations for new conditions, reducing deployment costs and expanding applicability.

Unified Multiple Condition Synthesis. Our method addresses a key limitation of existing approaches by enabling seamless integration of heterogeneous conditional signals. Tea-Adapter uses condition-specific pathway routing and adaptive fusion to resolve potential conflicts, aggregating features from diverse modalities (e.g., sketch, trajectory, normal map) while maintaining coherence. As shown in Figure 8, combining sketch (structure), trajectory (motion), and normal map (depth) yields outputs that satisfy all constraints without compromise. This unified control is valuable for animation production and interactive content creation, streamlining multi-dimensional control and enhancing creative flexibility.

6. Conclusion

This paper has proposed *Tea-Adapter*, a novel reversed distillation adapter for efficiently transferring conditional information from small teacher models to large student video diffusion transformers. Our method achieves robust performance in diverse video tasks with significantly reduced training costs. We plan to extend this architecture to other tasks, leveraging small models that are pre-trained on specific tasks and then transferring their capabilities to the foundational large model. This would allow the knowledge and capabilities of these task-specific models to be efficiently transferred and composed within a larger foundational video generation model. Such an extension could enable more controllable and semantically-aware synthesis across diverse conditional inputs, while keeping the training and inference costs tractable.

7. Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62502410) and the Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515011138).

References

[1] Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, et al. Ai flow: Perspectives, scenarios, and approaches. *Vicinagearth*, 3(1):1, 2026. 3

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[3] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Infinite-canvas: Higher-resolution video outpainting with extensive content generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2150–2158, 2025. 3

[4] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 3

[5] Yiyang Chen, Xuanhua He, Xiujun Ma, and Yue Ma. Contextflow: Training-free video object editing via adaptive context enrichment. *arXiv preprint arXiv:2509.17818*, 2025. 3

[6] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance. *arXiv preprint arXiv:2311.12886*, 2023. 3

[7] Karachev Denis. Dilated controlnet for wan2.1. *GitHub repository*, 2025. 3, 6

[8] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023. 3

[11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024. 3

[12] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[14] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2022. 3, 5

[15] 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. In *ICLR*, pages 1–26, 2022. 4

[16] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023. 3

[17] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV 2023*, 2023. 3

[18] Xuelong Li. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8708–8714, 2022. 3

[19] Xingyuan Li, Jinyuan Liu, Zhixin Chen, Yang Zou, Long Ma, Xin Fan, and Risheng Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, pages 270–288. Springer, 2024. 3

[20] Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, and Jinyuan Liu. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7534–7544, 2025.

[21] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[22] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024. 2, 3, 6

[23] Hongyu Liu, Xuan Wang, Ziyu Wan, Yue Ma, Jingye Chen, Yanbo Fan, Yujun Shen, Yibing Song, and Qifeng Chen. Avatarartist: Open-domain 4d avatarization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10758–10769, 2025. 3

[24] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 3

[25] Zeqian Long, Mingzhe Zheng, Kunyu Feng, Xinhua Zhang, Hongyu Liu, Harry Yang, Linfeng Zhang, Qifeng Chen, and Yue Ma. Follow-your-shape: Shape-aware image editing via trajectory-guided region control. *arXiv preprint arXiv:2508.08134*, 2025. 3

[26] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 3

[27] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[28] Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Xuanhua He, Chenyang Zhu, Hongyu Liu, Yingqing He, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025. 3

[29] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025. 3

[30] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025. 3

[31] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 3

[32] Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, et al. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*, 2025. 3

[33] Yue Ma, Zhikai Wang, Tianhao Ren, Mingzhe Zheng, Hongyu Liu, Jiayi Guo, Mark Fong, Yuxuan Xue, Zixiang Zhao, Konrad Schindler, et al. Fastvmt: Eliminating redundancy in video motion transfer. *arXiv preprint arXiv:2602.05551*, 2026. 3

[34] 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. In *AAAI 2024*, 2023. 2, 3

[35] OpenAI. Video generation models as world simulators. 2024. 3

[36] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2023. 2, 3

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. 2

[38] 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. In *NeurIPS 2023*, 2023. 3, 6

[39] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 3

[40] 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. 3

[41] Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8775–8784, 2024. 6

[42] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2, 3

[43] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 5

[44] Yutao Shen, Junkun Yuan, Toru Aonishi, Hideki Nakayama, and Yue Ma. Follow-your-preference: Towards preference-aligned image inpainting. *arXiv preprint arXiv:2509.23082*, 2025. 3

[45] 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. In *ICLR*, 2023. 3

[46] Yiren Song, Danze Chen, and Mike Zheng Shou. Layertracer: Cognitive-aligned layered svg synthesis via diffusion transformer. *arXiv preprint arXiv:2502.01105*, 2025.

[47] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572*, 2025. 3

[48] Shikun Sun, Min Zhou, Zixuan Wang, Xubin Li, Tiezheng Ge, Zijie Ye, Xiaoyu Qin, Junliang Xing, Bo Zheng, and Jia Jia. Minimal impact controlnet: Advancing multi-controlnet integration. *arXiv preprint arXiv:2506.01672*, 2025. 2, 3

[49] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5

[50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[51] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao

Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 5

[52] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *Advances in Neural Information Processing Systems*, 37:96541–96565, 2024.

[53] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024. 3

[54] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 5

[55] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[57] Zexuan Yan, Yue Ma, Chang Zou, Wenteng Chen, Qifeng Chen, and Linfeng Zhang. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17474–17484, 2025. 3

[58] Beiyuan Zhang, Yue Ma, Chunlei Fu, Xinyang Song, Zhenan Sun, and Ziqiang Li. Follow-your-multipose: Tuning-free multi-character text-to-video generation via pose guidance. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3

[59] 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

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[61] 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. 3

[62] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025. 3

[63] 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. *Advances in Neural Information Processing Systems*, 2023. 3

[64] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 6