Tea-Adapter: Teacher Adapter for Efficient Conditional Generation

Supplementary Material

1. Latent Distribution Statistic Analysis

To evaluate the transferability of feature distributions between the large and small models, we randomly selected a single data sample as input and conducted a statistical analysis of the latent spaces in the Diffusion Transformer Blocks of both models. We pair the DiT Blocks of the 1.3B model and the 14B model according to a specific sequence (Wan2.1 14B-T2V block index: [0, 4, 8, 12, 16, 20, 24, 28, 32, 36]; Wan2.1-1.3B(fine-tuned from 1.3B-T2V) block index: [0, 2, 5, 8, 11, 14, 15, 18, 21, 24]), and then conduct statistical analysis.

The analysis metrics were divided into two components: (1) Post-PCA dimensionality reduction: Statistical measures including mean correlation, standard deviation correlation, covariance similarity, Wasserstein distance, etc.; (2) Calculation of direct distribution distance: Metrics calculated without dimensionality reduction, such as Wasserstein distance and matrix similarity.

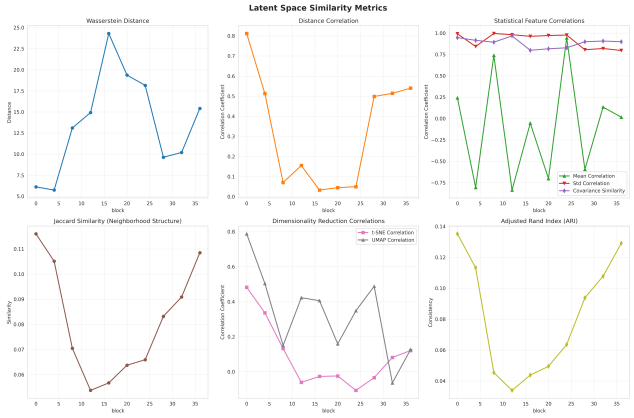

Figure 9. **Latent feature similarity between two scale models**. Observing the latent features of the first few and last few DiT Blocks in the two models exhibits high similarity, whereas the latent features of the middle Blocks show relatively low similarity. The two models exhibit volatility differences at different network levels, but the ultimately generated latent space representations have global consistency.

Results from this analysis are presented in Figure 9. We first analyzed the metrics from the paired DiT Blocks to characterize the relationship between the two latent space distributions. While distributional similarity varied substantially across blocks (e.g., Wasserstein distance ranged from 5.7 to 24.3, and distance correlation values spanned 0.03 to 0.81), classifier accuracy remained consistently close to 0.5 (range: 0.458–0.472) across all blocks. This observation indicates that although the local statistical properties of the two latent spaces differ, their overall representations are highly analogous and indistinguishable by a classifier across all model blocks. From our analysis results, we observe that the latent features of the first few and last few DiT Blocks in the two models exhibit high similarity, whereas the latent features of the middle Blocks show relatively low similarity. Inspired by this phenomenon, the Adapters in Tea-Adapter are primarily concentrated in the first few and last few Blocks, while the middle Blocks are equipped with fewer Adapters that are evenly spaced. This design is intended to effectively facilitate the transfer of features from the small model to the large model.

In Table 5, we compared the performance of early, deep, and uniform layer selections. Ultimately, we found that the uniform-layer setup achieved the best performance. Meanwhile, the analysis of Figure 11(e) reveals that early layers provide rich semantic information, while deep layers offer low-level features.

In conclusion, our analysis of the paired DiT Blocks reveals that the latent space distributions of the two models are generally analogous, though their similarity varies substantially across blocks: some blocks exhibit strong similarity, while others show notable divergence. For the feature transfer, a block-by-block strategy is therefore well-suited blocks with high similarity metrics are more amenable to direct transfer or distillation. Building on this insight, our work aims to develop a method for adapting features across DiT Blocks between small and large video diffusion models to enhance the efficacy of reversed distillation.

Layer Selection	FID ↓	LPIPS ↓	SSIM ↑	CLIP ↑
Early layers	223.7	0.522	0.481	0.653
Deep layers	346.5	0.795	0.266	0.521
Uniform layers	**62.4**	**0.241**	**0.828**	**0.921**
w/o shared expert	193.9	0.443	0.675	0.761

Table 5. Ablation study of layer selection strategies.

2. Additional Architecture Detail

As shown in Figure 10, the core architecture of Tea-Adapter consists of three main components: an attention module, a Mixture of Condition Experts (MCE) layer, and a Feature Propagation Module.

The attention module is a fundamental component within each transformer block, comprising three key sub-modules: LayerNorm layers, a self-attention mechanism, and a cross-attention mechanism. The LayerNorm layer first normalizes the input features to stabilize training. The self-attention

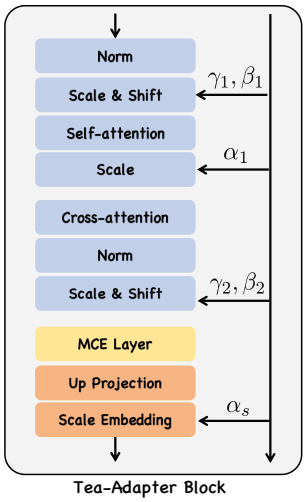

Figure 10. **Adapter detail**. Our design builds upon the DiT Block by incorporating the MCE Layer and an efficient Feature Propagation Module (Up Projection and Scale Embedding), which are the two key components enabling multi-condition integration and low-resource efficient training.

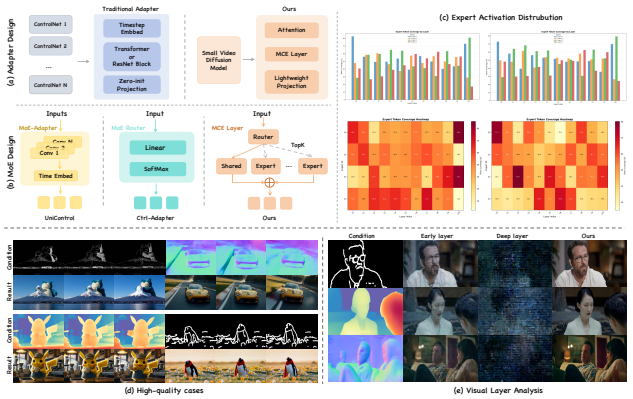

Figure 11. **(a)** Comparison of traditional adapter and our design. **(b)** Detail of different MoE variants in the adapter design. **(c)** Activation analysis of experts in different layers. **(d)** More high-quality visual cases. **(e)** Visual ablation analysis results in different layers.

mechanism then captures contextual dependencies among spatial and temporal tokens. Finally, the cross-attention module integrates conditional information (such as text or structural guidance) into the visual representation. This design enables effective fusion of spatial, temporal, and conditional features throughout the diffusion process.

The MCE layer includes the router, experts, and shared experts. Each expert is made up of an MLP layer. The default setting has 1 shared expert and 3 task-specific experts, and the topk weight is 2.

The Feature Propagation Module consists of a scale embedding layer and an up-projection layer. The scale embedding layer includes a scale factor α_s and a time embedding layer that can efficiently achieve the transmission of control condition information. The time embeddings of other layers are all initialized from the small model. These low-dimensional features encapsulate condition-specific information and enable seamless integration with the pre-trained small model. To support efficient training and inference, the MCE layer can be optionally removed, reducing the total parameter count by up to 50% without compromising performance. The Feature Propagation Module comprises an up-projection layer and a scale embedding mechanism, which together facilitate efficient and robust transfer of conditional features to the large diffusion model.

3. Training Setting and Data Process

Our model is trained with the following key hyperparameters: a learning rate of 2×10^{-5}, a training batch size of 2, a video sampling strategy that selects 24 frames per video with a sampling stride of 2, a video sample resolution of 512×512 (denoted as video sample size=512), a token sequence length of 512 (denoted as token sample size=512), 1 gradient accumulation step to stabilize training, and mixed-precision training using the "bf16" (bfloat16) format to balance computational efficiency and numerical precision [2, 4, 5, 15].

For condition data preparation, we leverage three conditions extracted from input frames: depth maps obtained via Depth-Anything (a monocular depth estimation model), structural boundaries extracted using a Canny edge detector, and human pose skeletons retrieved through the OpenPose framework, ensuring the model captures both global scene geometry and fine-grained semantic details [1, 6–13, 16**?** , 17].

4. Zero-Shot with Unseen Conditions

Our method demonstrated strong zero-shot generalization ability after a small number of conditional adaptation training. Our model is trained exclusively on conditional video data, including depth, pose, and Canny conditions. To evaluate its zero-shot generalization capability, we test the model on several previously unseen conditional inputs, including normal maps, scribbles, segmentation maps, MLSD edges, and line art. The result are shown in Table 6

Zero-shot performance analysis. The zero-shot generalization ability of our method stems primarily from the shared experts in the MCE layer, which capture common features across different conditions. This is supported by the ablation study in Table 5, where removing the shared expert leads to a rapid decline in performance. However, it still faces challenges in certain complex scenarios, such as videos with complex human motions or highly detailed visual content. We will improve in future work.

Table 6. Zero-shot Generation Performance on Unseen Conditional Inputs

Conditional	FVD (↓)	LPIPS (↓)	SSIM (↑)	CLIP (↑)
Canny(trained)	289.564	0.25	0.59	0.92
Normal Map	383.858	0.26	0.51	0.86
Scribble	350.504	0.25	0.52	0.87
Segmentation Map	385.048	0.27	0.49	0.86
MLSD	393.302	0.25	0.51	0.86
Line Art	331.156	0.23	0.52	0.88

Table 7. Experiment Between Different Scale Models

Model	FVD (↓)	CLIP (↑)	LPIPS (↓)	SSIM (↑)
Wan-Control-1.3B(T2V)	563.818	0.785	0.615	0.337
Wan-Control-14B(T2V)	**501.102**	**0.801**	**0.576**	**0.351**
Ours(T2V)	558.492	0.786	0.629	0.341
Wan-Control-1.3B(I2V)	312.613	0.896	0.205	0.572
Wan-Control-14B(I2V)	**254.238**	0.913	**0.193**	**0.664**
Ours(I2V)	303.202	**0.914**	0.218	0.584

Interpretability of MCE. In Figure 11(c), first row, we analyze expert activation across MCE layers under identical settings but varying conditions. The highly similar yet distinct activation patterns indicate that experts capture shared representations across conditions. The second row shows that applying different conditions activates distinct experts in each layer, suggesting specialization, where each expert handles a specific type of knowledge.

5. Additional Experiment

Our method is also competitive compared to models that have undergone large-scale pre-training under the same architecture. It is noted that both our large and small models are initialized from text-to-video models, with training data of less than 100k and a time of less than 48 GPU hours. We have conducted additional experiments comparing the performance of our method against 1.3B and 14B parameter models on both Image-to-Video (I2V) and Text-to-Video (T2V) generation tasks. For the baseline models, we used the pre-trained Wan2.1-Fun-Control model with 1.3B and 14B parameters, respectively. Our approach is based on the Wan2.1-14B-T2V model, with depth maps consistently applied as the control condition. The I2V test set comprises 100 samples selected from the Koala-36M dataset [14], whereas the T2V evaluation utilizes 1,000 samples from the Panda-70M dataset [3].

As shown in Table 7, experimental results demonstrate that our method achieves comparable performance to the heavily trained 1.3B and 14B models in the I2V task, and even outperforms the 14B model in the T2V task, demonstrating highly competitive generation quality. These results confirm that our approach effectively transfers knowledge from the small conditional model to the large base model, achieving strong performance with greater parameter efficiency.

References

[1] Kaitong Cai, Jusheng Zhang, Yijia Fan, Jing Yang, and Keze Wang. Racot: Plug-and-play contrastive example generation mechanism for enhanced llm reasoning reliability, 2025. 3

[2] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 3

[3] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 4

[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] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240): 1–113, 2023. 3

[6] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 3

[7] Yijia Fan, Jusheng Zhang, Kaitong Cai, Jing Yang, Chengpei Tang, Jian Wang, and Keze Wang. Cost-effective communication: An auction-based method for language agent interaction, 2025.

[8] 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. In *ICLR*, 2023.

[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.

[10] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2017.

[11] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[12] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

[13] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. 2022. 3

[14] 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. 4

[15] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 3

[16] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm:counterfactual vision-language fine-tuning, 2025. 3

[17] Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. GAM-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3