

SUPPLEMENTARY MATERIALS

Anonymous authors

Paper under double-blind review

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. Results from this

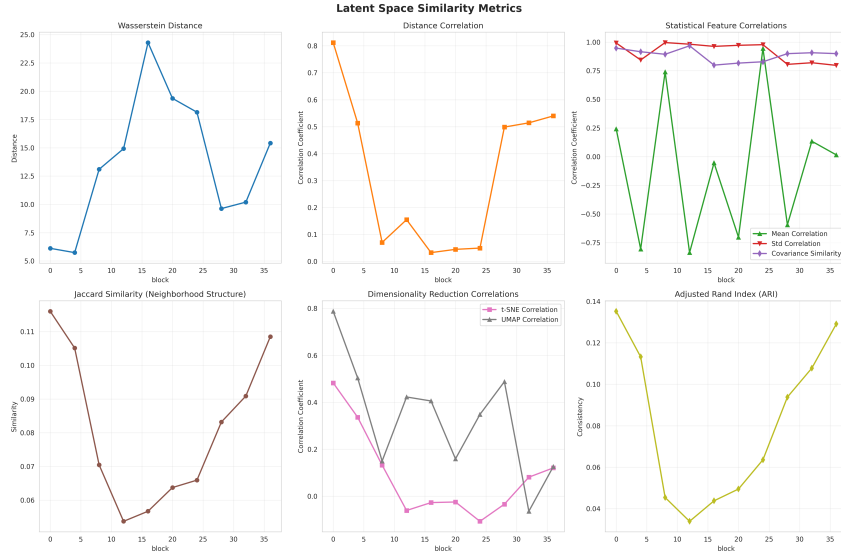


Figure 9: Latent feature similarity between two scale models. Observing the latent features of the first few and last few DiT Blocks in the two models exhibit high similarity, whereas the latent features of the middle Blocks show relatively low similarity. The two models exhibit volatility differences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

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 Scale-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 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 task, 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.

2 ADDITIONAL ARCHITECTURE DETAIL

As shown in Figure 10, the core architecture of Scale-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 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.

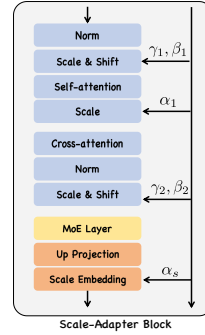


Figure 10: **Adapter detail.** Two important principles of design are multi-condition integration and low-resource efficient training.

3 TRAINING SETTING AND DATA PROCESS

4 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.

For condition data preparation, we leverage three conditions extracted from input frames: depth maps obtained via Depth-Anything (a state-of-the-art monocular depth estimation model), structural

Table 3: Zero-shot Generation Performance on Unseen Conditional Inputs

Conditional	FVD (\downarrow)	LPIPS (\downarrow)	SSIM (\uparrow)	CLIP (\uparrow)
Canny(trained)	1447.82	0.25	0.59	0.92
Normal Map	1919.29	0.26	0.51	0.86
Scribble	1752.51	0.25	0.52	0.87
Segmentation Map	1925.24	0.27	0.49	0.86
MLSD	1966.51	0.25	0.51	0.86
Line Art	1655.78	0.23	0.52	0.88

Table 4: Experiment Between Different Scale Models

Model	FVD (\downarrow)	CLIP (\uparrow)	LPIPS (\downarrow)	SSIM (\uparrow)
Wan-Control-1.3B(T2V)	2819.091	0.785	0.615	0.337
Wan-Control-14B(T2V)	2505.551	0.801	0.576	0.351
Ours(T2V)	2797.460	0.786	0.629	0.341
Wan-Control-1.3B(I2V)	1563.066	0.896	0.205	0.572
Wan-Control-14B(I2V)	1271.194	0.913	0.193	0.664
Ours(I2V)	1516.011	0.914	0.218	0.584

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.

5 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.

Quantitative results are summarized in Table 3, where our method is compared against existing approaches using widely adopted metrics including FVD, LPIPS, SSIM, and CLIP score. The results demonstrate that our approach effectively adapts to novel conditions without additional fine-tuning, maintaining high visual quality and semantic alignment across diverse control signals. The experiment results demonstrate that our MCE layer not only supports multi-condition adaptation but also zero-shot learning.

6 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 (Wang et al., 2024), whereas the T2V evaluation utilizes 1,000 samples from the Panda-70M dataset (Chen et al., 2024).

As shown in Table 4, 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.

7 ETHICS STATEMENT AND REPRODUCIBILITY

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee. This article does not contain any studies with human participants performed by any of the authors. Informed consent was obtained from all individual participants included in the study. We will release code under **CC-BY-NC-4.0**.

8 THE USAGE OF LARGE LANGUAGE MODELS

In this paper, the usage of the LLM mainly falls into the following aspects: specifically, for grammar checking and format optimization, we use DeepSeek-R1 to conduct grammar error checking on the paragraphs of the paper as well as format checking of charts and graphs; additionally, for language polishing, we apply Doubao to polish and optimize the language expression of the paper’s text description part; and it is important to note that all authors are responsible for the content generated by the LLMs.

REFERENCES

- 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, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers, 2024. URL <https://arxiv.org/abs/2402.19479>.
- Qiuhe Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content, 2024. URL <https://arxiv.org/abs/2410.08260>.