EFFICIENT CONDITIONAL MODELING FOR CONTROLLABLE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

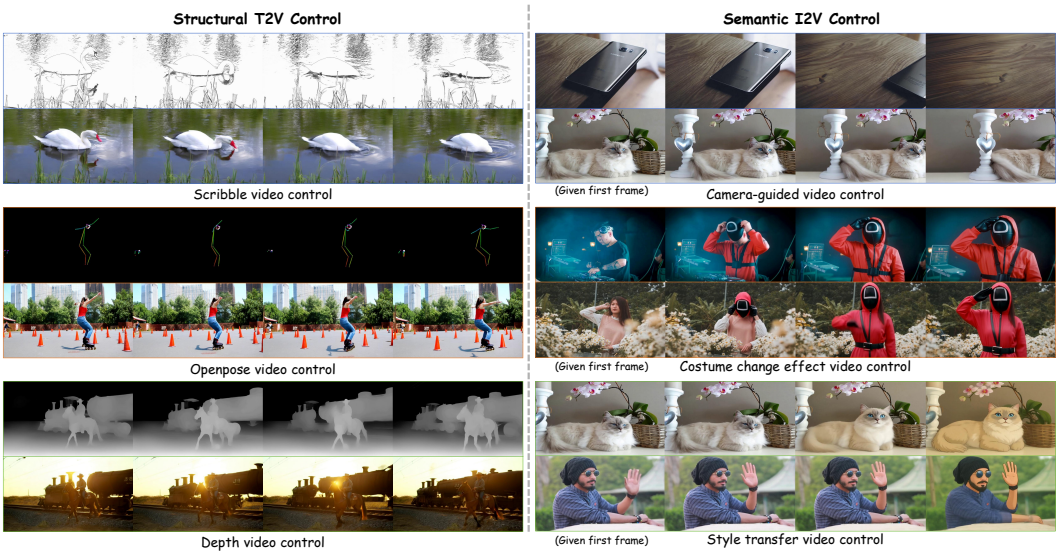

Figure 1: We introduce **EasyVideoControl**, a lightweight, plug-and-play conditional module for video diffusion transformers. It unifies structural (pose, edges) and semantic (camera- or style-guided) conditioning for text-to-video and image-to-video generation. EasyVideoControl trains with only 10% trainable parameters, reduces the overhead of naive conditioning mechanisms, and introduces only 5–8% inference overhead compared to standard text-to-video (16.9s $\rightarrow$ 17.7s) and image-to-video (18.0s $\rightarrow$ 19.4s) pipelines, measured on 832$\times$480 49-frame samples on a single H800 GPU, achieving universal control with minimal additional cost.

ABSTRACT

Despite substantial recent progress in controllable video generation with various conditioning signals, current frameworks suffer from severe computational inefficiency when applied to video diffusion transformers. The issue becomes particularly acute when the conditioning signal itself is a video, resulting in excessively long contextual sequences. To address this limitation, we propose EasyVideoControl, an efficient framework for conditional video generation with DiTs (diffusion transformers). EasyVideoControl introduces two core components. First, we present the Asymmetric Shared Causal Mixture-of-Transformers structure, which incorporates a frozen copy of the base diffusion transformer that shares self-attention layers with the backbone under a causal attention scheme. By training only low-rank adaptation (LoRA) parameters in this unidirectional conditioning branch, the model maximizes reuse of pretrained knowledge while minimizing training computational overhead. This design enables the backbone to be modulated by the conditional branch without reciprocal dependency, supporting KV Cache for efficient inference. Second, we introduce Reconstructive Conditioning Compression (RCC), a lightweight encoderdecoder mechanism that compresses and reconstructs conditional tokens. This approach greatly reduces context length while preserving fine-grained conditional details through a reconstruction-based objective. Extensive experiments across multiple conditional video generation tasks show that EasyVideoControl achieves comparable or superior performance while being significantly more efficient than existing video control methods.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent diffusion-based text-to-video models Chen et al. (2023a; 2024); Hong et al. (2022); Yang et al. (2024); Kong et al. (2025); Wan et al. (2025); Lin et al. (2024); Ho et al. (2022); Blattmann et al. (2023) have demonstrated remarkable progress in synthesizing high-fidelity, semantically coherent videos. This progress has also sparked increasing interest in *controllable video generation*, which aims to guide temporal content through explicit structural signals (such as poses or depths) or semantic signals (including camera motion, style, or visual effects). Controllable generation enables greater consistency, editability, and alignment with user intent, making it critical for applications in film production, virtual avatars, and so on.

Conditioning mechanisms have evolved alongside diffusion model architectures. Under U-Net backbones, ControlNet-style methods Zhang et al. (2023); Chen et al. (2023b); Peng et al. (2024); Guo et al. (2024); Zhang et al. (2024b) inject structural control through residual feature injection, while others Fang et al. (2024); Pang et al. (2024); Wei et al. (2024); Jiang et al. (2024); Wu et al. (2024) add cross-attention layers for semantic reference inputs. With the shift from U-Nets to diffusion transformers (DiTs), token concatenation has become the practical conditioning paradigm Ye et al. (2025); Luo et al. (2025); Wang et al. (2025b); Zhang et al. (2025b); Wang et al. (2025a); Huang et al. (2025): conditional tokens are appended to visual latent tokens and jointly processed by all transformer layers, offering a unified interface for both structural and semantic signals. Throughout this work, semantic controls (e.g., camera motion or style) are supplied as full-length reference videos following the Video-as-Prompt paradigm Bian et al. (2025a) rather than low-dimensional parametric sequences such as camera extrinsics, and thus produce dense spatiotemporal tokens just as structural conditions do.

This flexibility, however, comes at a severe computational cost, as attention scales quadratically ($O(N^2)$) with the number of latent tokens N, which grows with both frame count and spatial resolution. The overhead arises at two levels: (i) *intra-step context bloat*, where appending condition tokens to the already long video sequence inflates the compute and activation memory of every attention evaluation; and (ii) *inter-step redundancy*, where condition tokens—entangled with the noisy latents and timestep-dependent modulation—are re-encoded and re-attended at every denoising step although the condition itself is static. EasyVideoControl is designed around these two observations.

Concurrently, image-based DiT conditioning has seen efficiency-driven advances such as EasyControl Zhang et al. (2025a) and OminiControl2 Tan et al. (2025). These 2D mechanisms, however, do not transfer directly to video, where spatiotemporal token sequences are orders of magnitude longer: caching their uncompressed key–value states across all transformer layers strains GPU memory, and naively downsampling the spatiotemporal grid breaks the correspondence with the backbone's 3D positional encodings, causing spatial drift unless positions are explicitly realigned. To bridge this gap, we propose EasyVideoControl, which re-optimizes these efficiency principles for the high-dimensional video manifold, enabling scalable, fine-grained, and computationally efficient conditional modeling in video diffusion transformers.

At the architectural level, we introduce the *Asymmetric Shared Causal Mixture-of-Transformers* (ASC-MoT), where a conditional branch shares self-attention layers with the backbone under a causal dependency scheme. Conditional tokens influence the main denoising stream in a unidirectional manner, reusing pretrained backbone knowledge while updating only lightweight LoRA parameters. This asymmetric design enables efficient KV caching during inference and substantially reduces redundant transformer computations without sacrificing control fidelity.

Complementary to the architecture, we propose *Reconstructive Conditioning Compression* (RCC), a lightweight encoderdecoder that compresses conditional video tokens into a compact latent sequence while preserving structural and semantic details: structure-aware token grouping is aligned with the backbone's positional encodings, and a jointly trained reconstruction objective retains fine-grained conditional information. By compressing conditional tokens before injection, RCC substantially shortens the context, accelerating both training and inference.

In summary, our contributions are threefold:

- We propose an Asymmetric Shared Causal MoT architecture that reuses pretrained transformer attention for unidirectional conditioning, achieving efficient inference with KV

caching and parameter-efficient training via LoRA with only minimal additional parameters.

- We introduce a Reconstructive Conditioning Compression mechanism that compresses conditional tokens while retaining fine-grained structural and semantic information. The resulting reduction in sequence length further accelerates both training and inference.
- Extensive experiments across structural and semantic controllable video generation tasks demonstrate that EasyVideoControl achieves competitive or superior controllability and synthesis quality while being considerably more efficient than existing approaches.

2 Related Works

2.1 Video Diffusion Models.

Diffusion models Ho et al. (2020); Song et al. (2020); Rombach et al. (2022); Esser et al. (2024); Podell et al. (2023); Peebles & Xie (2023); Black Forest Labs (2024) have become the state of the art for image synthesis, and their extension to the video domain has enabled dynamic and temporally coherent generation Blattmann et al. (2023); Guo et al. (2023); Zhang et al. (2024a); Singer et al. (2022); Ho et al. (2022); Khachatryan et al. (2023); Wang et al. (2023a); Chen et al. (2023a; 2024). Early approaches augment pretrained image backbones with temporal modules Guo et al. (2023), while subsequent models improve temporal consistency and fidelity through spatiotemporal training Chen et al. (2023a; 2024); Blattmann et al. (2023). More recently, Diffusion Transformers (DiT) Peebles & Xie (2023) have replaced the commonly used U-Net backbones with a more powerful and flexible architecture. Following this trend, recent open-source DiT-based video diffusion models, including Open-sora-plan Lin et al. (2024), CogVideoX Hong et al. (2022); Yang et al. (2024), Wan2.2 Wan et al. (2025), and HunyuanVideo Kong et al. (2025), adopt spatiotemporal VAEs and DiT structures for large-scale training.

2.2 Controllable Video Generation

Controllable video generation aims to guide synthesis through external conditioning signals. Under U-Net backbones, methods following the ControlNet paradigm Zhang et al. (2023) inject features of an auxiliary branch into the diffusion backbone via residual connections Chen et al. (2023b); Peng et al. (2024); Zhang et al. (2024b); Wang et al. (2024), focusing mainly on *structural control* signals such as poses, depth maps, or edge maps. With diffusion transformers, conditioning has converged to two lines: full-framework adaptation, as in VACE Jiang et al. (2025), which processes multiple control modalities through a Video Condition Unit and a context adapter at the cost of fine-tuning the full backbone; and the *token concatenation* paradigm, where conditional features are encoded as tokens and concatenated with noisy latent tokens at the input of the transformer. The latter flexibly accommodates not only structural controls but also diverse *semantic controls*—including reference-based style transfer Ye et al. (2025); Bian et al. (2025a); Li et al. (2025), camera control Ye et al. (2025); Bian et al. (2025a); Luo et al. (2025); Li et al. (2025), personalized visual effects Bian et al. (2025a); Li et al. (2025), identity conditioning Ye et al. (2025), and scene control Yu et al. (2025)—in which guidance is provided by dense reference videos rather than low-dimensional parametric signals. Domain-specific controllers instead target camera trajectories He et al. (2024); Wang et al. (2023c) or pose-guided human animation Hu et al. (2023); Zhang et al. (2024c) with dedicated designs, whereas our goal is efficient dense-video conditioning within a unified framework. However, encoding video conditions produces extremely long token sequences, and jointly processing them incurs prohibitive memory consumption and quadratic attention cost, hindering the scalability and practicality of controllable video diffusion transformers in real-world settings.

3 Method

In this section, we present EasyVideoControl, a framework for efficient conditioning modeling in video diffusion transformers. As illustrated in Fig. 2, our framework is built upon two synergistic components. (1) The Asymmetric Shared Causal Mixture-of-Transformers (ASC-MoT) (Sec. 3.1), which enables parameter-efficient conditional video diffusion and fast inference by caching the condition token KV states instead of recomputing them at every diffusion step. (2) The Reconstructive

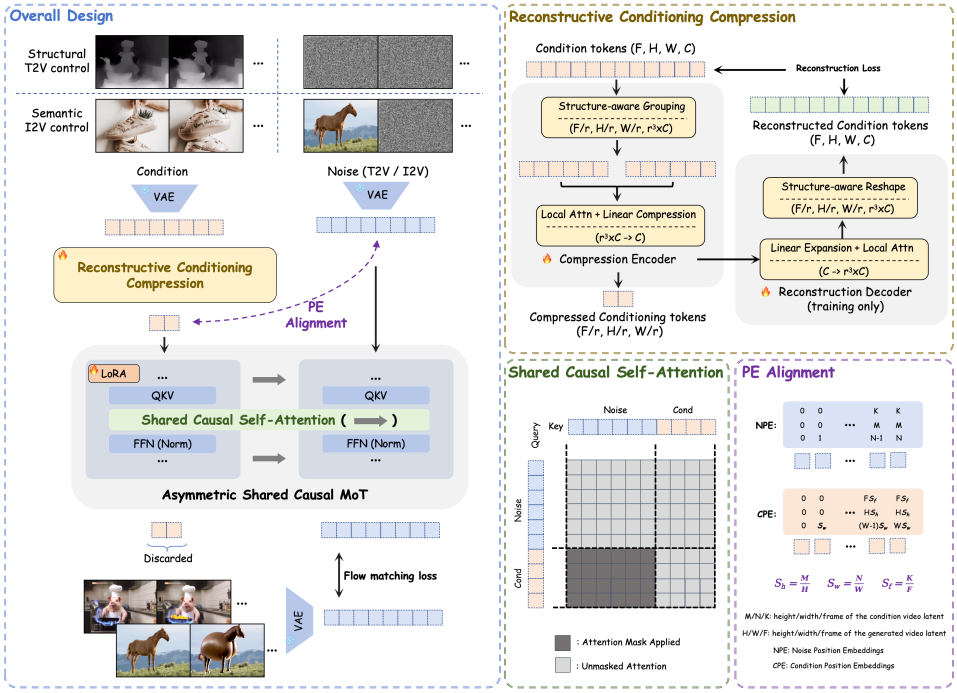

Figure 2: **Architecture of *EasyVideoControl*.** The framework integrates two cooperating modules: (1) the *ASC-MoT* module, which introduces a causal conditioning branch with shared self-attention and parameter-efficient training, and enables KV caching by preventing condition tokens from being re-encoded or jointly re-attended with video tokens at every diffusion step; and (2) the *Reconstructive Conditioning Compression* module, which compresses long spatiotemporal condition sequences via structure-aware grouping and PE alignment, preserving geometric fidelity while reducing token length by up to $8\times$. The combined design yields efficient and high-quality controllable video generation.

Conditioning Compression (RCC) (Sec. 3.2), which performs structure-aware spatiotemporal token compression, achieving up to $8\times$ reduction in conditional token length, thereby significantly improving both training and inference efficiency. Together, these components enable efficient and expressive video conditioning with minimal compute and memory overhead.

3.1 Asymmetric Shared Causal MoT

A key design objective of EasyVideoControl is to separate condition encoding from per-step denoising, thereby avoiding repeated token concatenation and enabling KV cachebased acceleration during sampling. To achieve this, we introduce an *Asymmetric Shared Causal Mixture-of-Transformers (ASC-MoT)* architecture. Instead of naively concatenating conditional tokens with noisy video tokens and jointly processing them through all transformer blocks, ASC-MoT augments the base DiT with an auxiliary *conditional branch*. This branch operates in parallel with the main denoising branch and maintains an explicit causal dependency: conditional tokens influence the denoising process, but not vice versa.

3.1.1 Architectural Design.

Formally, let f_θ be the pretrained DiT denoiser producing video tokens. Instead of training a fully learnable second DiT as in prior MoT-style conditioning Bian et al. (2025a), we instantiate a frozen structural copy of f_θ with lightweight LoRA adapters, $f_\theta^{\text{cond}} = f_\theta^{\text{frozen}} + \Delta_\phi^{\text{LoRA}}$, where ϕ denotes a small set of trainable parameters. This configuration drastically reduces trainable parameters and inherits pretrained inductive priors for consistent feature alignment with the backbone.

Each ASC-MoT block shares a self-attention layer between the backbone and conditional branches, while keeping the feed-forward networks (FFNs) disjoint. Let $X_B \in \mathbb{R}^{N_v \times d}$ and $X_C \in \mathbb{R}^{N_c \times d}$ denote backbone and conditional tokens, respectively. We form the combined sequence $X = [X_B; X_C] \in \mathbb{R}^{(N_v+N_c)\times d}$ and apply the shared self-attention module:

$$[Y_B; Y_C] = \text{SA}_{\text{shared}}(\text{norm}(X)) + X.$$

The backbone applies its frozen FFN, while the conditional branch uses a LoRA-augmented FFN:

$$Z_B = \mathrm{FFN}_B(\mathrm{norm}(Y_B)) + Y_B, \qquad Z_C = (\mathrm{FFN}_C + \Delta W)\,(\mathrm{norm}(Y_C)) + Y_C,$$

where $\Delta W = BA$ is a rank-r update ($r \ll d$) and only A and B are trainable.

3.1.2 Shared Causal Self-Attention.

The computational efficiency of ASC-MoT during inference is a direct consequence of a specially formulated attention mask, M, applied within the shared self-attention module, SA_{shared}. The standard scaled dot-product attention (with d_k the per-head attention dimension) is modified to:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{1}$$

The mask $M \in \mathbb{R}^{(N_v+N_c)\times(N_v+N_c)}$ is a block matrix that enforces a unidirectional information flow. It is defined as:

$$M_{ij} = \begin{cases} 0 & \text{if } i \le N_v \\ 0 & \text{if } i > N_v \text{ and } j > N_v \\ -\infty & \text{if } i > N_v \text{ and } j \le N_v \end{cases} \tag{2}$$

This mask allows the backbone tokens ($i \le N_v$) to attend to all tokens, thus incorporating the conditional guidance. However, it prevents the condition tokens ($i > N_v$) from attending to the backbone tokens ($j \le N_v$). This causal separation ensures that the computation of the conditional branch is independent of the noisy input z_t, allowing its representation to be computed once and cached during the iterative denoising process.

3.1.3 Inference Efficiency via KV Caching.

The asymmetric causal formulation has a critical inference benefit: conditional KV states can be computed once and reused across the denoising trajectory. To make this valid, we set the timestep modulation of condition tokens to zero, i.e., no t-dependent modulation is applied to the conditional branch. Combined with the causal attention mask above, this makes the conditional branch timestep-independent during both training and sampling. Specifically, we compute: $(K_c, V_c) = f_\theta^{\text{cond}}(\mathbf{c}, t{=}0)$ once per control video, and treat (K_c, V_c) as cached condition states that augment each denoising block. Since conditional tokens are only read by the video denoising tokens and are never updated by video latents, their key–value states remain valid across all diffusion steps. As a result, conditional features are computed once and reused throughout sampling, eliminating redundant transformer passes and yielding substantial speedups.

3.2 Reconstructive Conditioning Compression

Many video-based conditions produce long token sequences L_c, which leads to significant overhead during attention computation. We introduce a *Reconstructive Conditioning Compression (RCC)* module designed to (1) drastically reduce the number of conditional tokens, while (2) preserving structural and semantic fidelity necessary for controllable video generation.

3.2.1 Structure-Aware Token Compression.

Let $c_{\text{orig}} \in \mathbb{R}^{L_c \times d_c}$ denote the original conditional token sequence obtained from a spatiotemporal encoder. To reduce token length while preserving structural fidelity, we introduce a structure-aware compression encoder $\mathcal{E}(\cdot)$ that performs compression along temporal and spatial dimensions prior to linear projection.

Concretely, conditional tokens are first reshaped into (F, H, W, C) and partitioned into non-overlapping spatiotemporal groups of size (r_t, r_h, r_w). Within each group, we apply a lightweight local self-attention to aggregate neighborhood information, followed by a linear projection that reduces channel dimensionality. This yields

$$c_{\text{comp}} = \mathcal{E}(c_{\text{orig}}) \in \mathbb{R}^{(F/r_t \cdot H/r_h \cdot W/r_w)\times d_c'}, \qquad d_c' \le d_c, \tag{3}$$

achieving a controllable compression ratio of $r = r_t r_h r_w$ while maintaining temporal consistency and spatial continuity. The resulting sequence length $N_c = L_c/r$ is the length of the conditional



Figure 3: **Qualitative Comparison.** Compared to existing SOTA baselines, EasyVideoControl has superior performance in maintaining fine-grained structural fidelity and temporal stability, while requiring fewer conditional tokens and significantly lower computational overhead.

tokens X_C used by ASC-MoT in Sec. 3.1. In our implementation, we strike a balance between efficiency and fidelity by setting a uniform compression factor, *i.e.*, $r_t = r_h = r_w = 2$, resulting in an $8\times$ reduction in token sequence length.

3.2.2 Position Encoding Alignment.

Compression changes the spatial–temporal resolution of conditional signals, potentially breaking positional correspondence with DiT video tokens. To preserve geometric consistency, we apply *Position Encoding Alignment (PEA)*, which rescales positional encodings according to temporal and spatial compression ratios.

Let the original condition grid be (F, H, W) and the compressed grid be (F', H', W'). We define

$$S_t = \frac{F}{F'}, \quad S_h = \frac{H}{H'}, \quad S_w = \frac{W}{W'}. \tag{4}$$

For each compressed token (t, i, j), its aligned position is

$$(P_t, P_i, P_j) = (tS_t,\ iS_h,\ jS_w). \tag{5}$$

This preserves spatiotemporal correspondence between compressed condition tokens and video tokens, enabling consistent controllable generation under aggressive token reduction.

3.2.3 Token Reconstruction.

To ensure that c_{comp} remains semantically and spatially sufficient for control, we train a symmetric decoder $\mathcal{D}(\cdot)$ to reconstruct the original tokens:

$$\hat{c}_{\text{orig}} = \mathcal{D}(c_{\text{comp}}).$$

The encoder–decoder pair $(\mathcal{E}, \mathcal{D})$ is trained jointly with the main diffusion model using a reconstruction objective

$$\mathcal{L}_{\text{rec}} = \|c_{\text{orig}} - \mathcal{D}(\mathcal{E}(c_{\text{orig}}))\|_2^2. \tag{6}$$

The total training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fm}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}, \tag{7}$$

where $\mathcal{L}_{\text{fm}}$ is the flow matching loss. In our experiments, we set the balancing hyperparameter $\lambda_{\text{rec}} = 0.01$.

Type	Method	Video Quality&Video Consistency							Controllability	Efficiency	
		Aesthetic Quality	Background Consistency	Dynamic Degree	Imaging Quality	Motion Smoothness	Overall Consistency	Temporal Flickering	MAE Error	Latency	Trainable Params Percentage
Structural T2V Control	Control-A-Video	50.62	91.71	70.00	67.76	97.58	24.48	96.58	0.024	95.9s	55%
	ControlVideo	**63.30**	95.02	10.00	65.13	96.49	24.20	95.42	0.093	333.4s	-
	ControlNet†	56.00	94.04	70.00	63.69	98.17	25.48	96.46	0.044	38.7s	7%
	VACE(Wan-1.3b)	61.04	95.90	70.00	**67.94**	97.91	25.40	96.31	0.026	51.0s	100%
	Ours(8x)	58.15	**96.24**	65.00	59.65	**98.67**	**25.51**	**97.08**	**0.022**	**17.7s**	10%
	Ours(8x_W/O_Rec)	59.56↑	96.37↑	60.00↓	60.50↑	98.60↓	25.95↑	96.94↓	0.028↓	17.7s	-
	Ours(8x_W/O_PE)	60.56↑	96.74↑	55.00↓	64.54↑	98.80↑	26.07↑	97.10↑	0.034↓	17.7s	-
	Ours(8x_W/O_Causal)	61.98↑	96.99↑	60.00↓	66.62↑	98.79↑	25.80↑	97.11↑	0.024↓	27.9s↓	-
	Ours(27x)	63.08↑	96.69↑	65.00	67.49↑	98.76↑	26.09↑	97.14↑	0.031↓	17.6s↑	-
	Ours(27x_W/O_Rec)	64.01↑	96.89↑	65.00	69.69↑	98.87↑	25.84↑	97.16↑	0.044↓	17.6s↑	-
	Ours(64x)	62.77↑	97.12↑	70.00↑	67.87↑	98.88↑	26.01↑	97.16↑	0.043↓	17.4s↑	-
	Ours(64x_W/O_Rec)	63.75↑	96.94↑	70.00↑	68.94↑	98.82↑	25.97↑	97.11↑	0.046↓	17.4s↑	-
	Ours(W/O_Compress)	59.14↑	96.35↑	70.00↑	62.46↑	98.61↓	25.60↑	96.91↓	0.020↑	20.7s↓	-
	Text2Video	-	-	-	-	-	-	-	-	16.9s	-
Semantic I2V Control	VAP	57.71	-	77.08	-	**98.59**	24.13	-	-	317.0s	100%
	Kling/Vidu	59.16	-	79.17	-	98.12	24.05	-	-	-	-
	Ours(8x)	**59.35**	-	**81.25**	-	98.54	**24.68**	-	-	**19.4s**	10%
	Ours(8x_W/O_Rec)	59.82↑	-	75.00↓	-	98.50↓	24.23↓	-	-	19.4s	-
	Ours(8x_W/O_PE)	58.87↓	-	79.16↓	-	98.49↓	24.09↓	-	-	19.4s	-
	Ours(8x_W/O_Causal)	58.74↓	-	87.50↑	-	99.27↑	24.78↑	-	-	28.0s↓	-
	Ours(27x)	59.45↑	-	81.25	-	98.44↓	24.62↓	-	-	19.1s↑	-
	Ours(27x_W/O_Rec)	59.23↓	-	79.16↓	-	98.42↓	24.40↓	-	-	19.1s↑	-
	Ours(64x)	58.97↓	-	81.25	-	98.43↓	24.52↓	-	-	19.0s↑	-
	Ours(64x_W/O_Rec)	58.47↓	-	79.16↓	-	98.56↑	24.42↓	-	-	19.0s↑	-
	Ours(W/O_Compress)	59.75↑	-	83.33↑	-	98.44↓	25.16↑	-	-	22.4s↓	-
	Image2Video	-	-	-	-	-	-	-	-	18.0s	-

Table 1: **Quantitative evaluation.** We compare **EasyVideoControl** against SOTA structural and semantic control methods, alongside various ablation baselines. **Metrics:** For *Video Quality & Consistency* (e.g., Aesthetic, Motion Smoothness), higher values indicate better performance. For *Controllability* (MAE Error) and *Efficiency* (Latency), lower values are preferred. In the ablation study, the markers ↑ and ↓ denote **better** and **worse** performance, respectively, relative to our full model (**Ours 8**×). For higher-is-better metrics, this aligns with larger/smaller values; for lower-is-better metrics (e.g., MAE and latency), it is reversed. **Notes:** † uses a skip-layer variant Karachev (2025) of ControlNet to reduce parameters (7%) by copying only a subset of blocks, rather than the full backbone Zhang et al. (2023). **Backbones:** all *Ours* variants use Wan2.2-TI2V-5B, ControlNet† uses Wan2.2, and VACE uses Wan2.1-1.3B; Control-A-Video and ControlVideo are U-Net-based methods on earlier backbones. A same-backbone comparison with VACE on Wan2.1-1.3B is provided in Table 2.

This encourages $\mathcal{E}$ to learn a compressed latent sufficient statistic of the condition that preserves control-relevant information. During the training and inference, the compressed tokens c_{comp} are used as the conditional tokens X_C for the ASC-MoT module, thereby reducing attention complexity while retaining effective conditioning signals.

4 Experiments

4.1 Implementation Details

Settings. We use the pretrained Wan2.2-TI2V-5B Wan et al. (2025) video diffusion transformer. For both structural text-to-video (T2V) control and semantic image-to-video (I2V) control settings, we resize videos to $480\times832\times49$ (HWF) and train on 8×H800 (80GB) GPUs with a per-GPU batch size of 1 and a learning rate of 5e-5, training over 80k training steps. During inference, we use the flow-matching scheduler with 50 sampling steps.

Data. For training, we use about 30K real-world videos from VP-data Bian et al. (2025b) for structural T2V tasks and VAP-data Bian et al. (2025a) for semantic I2V tasks. For structural T2V, depth, scribble, and pose annotations are pre-extracted from the videos. For evaluation, structural T2V is assessed using the VACE benchmark Jiang et al. (2025), while semantic I2V is evaluated on the VAP benchmark Bian et al. (2025a) within VAP-data.

4.2 Comparison Methods

For the structural T2V task, we compare our method against several representative methods, including Control-A-Video Chen et al. (2023b), ControlVideo Zhang et al. (2024b), community-released video variants of ControlNet Zhang et al. (2023); Karachev (2025) and VACE Jiang et al. (2025). For semantic I2V tasks, we include comparisons with VAP Bian et al. (2025a), as well as some commercial systems (e.g., VIDU Vidu Team (2025) and Kling Kling AI Team (2025)). All methods are evaluated using their official or author-recommended inference configurations. Unless otherwise specified, videos are generated at **49 frames** with a spatial resolution of 480×832 and **50 sampling steps**.

4.3 Evaluation Metrics

To comprehensively assess the performance of each framework, we adopt VBench Huang et al. (2024) as our primary evaluation tool, which provides a detailed evaluation suite for both text-to-video (T2V) and image-to-video (I2V) tasks. For the structural T2V task, we evaluate seven dimensions: motion smoothness, dynamic degree, aesthetic quality, temporal flickering, imaging quality, background consistency, and overall consistency. For the semantic I2V task, following VAP Bian et al. (2025a), we measure motion smoothness, dynamic degree, aesthetic quality, and overall consistency (using CLIP score Sun (2023); Radford et al. (2021) computed via ViCLIP Wang et al. (2023b)). Following prior works Guo et al. (2024); Chen et al. (2023b); Zhang et al. (2025a); Tan et al. (2024), we additionally compute the MAE between depth maps estimated by MiDaS Birkl et al. (2023) from the generated frames and those from the original videos, evaluating the structural consistency of generation under depth conditions. To evaluate efficiency, we report the inference latency on a single H800 GPU and the proportion of trainable parameters relative to the base model, where a lower ratio indicates higher parameter efficiency.

4.4 Experiment Results

4.4.1 Qualitative comparison.

We present qualitative comparisons with the SOTA baselines in Fig. 3. For structural T2V tasks, EasyVideoControl demonstrates markedly stronger spatial fidelity. ControlVideo and Control-A-Video commonly introduce background warping or visible degradation when characters undergo motion. ControlNet variants produce visually appealing results, yet the imposed structural conditions are not faithfully preserved, leading to inconsistencies in both pose and geometric layout. In contrast, EasyVideoControl maintains stable backgrounds, accurate depth alignment, and strong prompt adherence, achieving visual quality on par with VACE. Notably, our method attains these quality levels while operating with significantly lower computational cost, resulting in substantially faster inference in practice. For semantic I2V tasks, EasyVideoControl excels at preserving fine-grained patterns and maintaining layout structure during temporal transformations. For example, when converting a real craftsman into a Simpsons-style animation or transforming a person into a voxel block character, our method retains pose direction and spatial composition throughout the transition. In contrast, VAP frequently loses structural layout or produces semantic inconsistencies relative to the input video. These results demonstrate that EasyVideoControl can preserve high-frequency details and temporal coherence, while operating with significantly higher computational efficiency.

4.4.2 Quantitative comparison.

Table 1 presents a comprehensive quantitative comparison between EasyVideoControl and SOTA methods for both structural (T2V) and semantic (I2V) controllable generation. In T2V tasks, our full model (Ours $8\times$) achieves the best control precision with the lowest MAE Error, outperforming specialized models like VACE and ControlNet. Notably, while maintaining high fidelity, our method achieves a significant breakthrough in efficiency, reducing Latency to 17.7s, nearly a $3\times$ speedup compared to VACE (51.0s). We note that VACE is instantiated on the smaller Wan2.1-1.3B backbone; under the identical backbone (Table 2), EasyVideoControl attains comparable quality and controllability while reducing latency from 51.0s to 10.5s, indicating that the efficiency gains stem from the conditioning design rather than backbone capacity. In I2V tasks, our model demonstrates

superior performance in Aesthetic Quality (59.35) and Overall Consistency (24.68), surpassing established research baselines and commercial models such as VAP and Kling/Vidu. By utilizing parameter-efficient training, our framework requires only 10% trainable parameters, making it substantially more efficient than other approaches like VACE or VAP while delivering competitive or superior visual quality. Notably, our method introduces only 5% and 8% inference overhead compared to standard T2V (16.9s $\rightarrow$ 17.7s) and I2V (18.0s $\rightarrow$ 19.4s) pipelines.

4.4.3 Ablation Study.

Tab. 1 reports ablations on the two core components of EasyVideoControl: RCC and ASC-MoT. Removing the reconstruction objective (*W/O_Rec*) consistently degrades controllability for both structural setting, reflected by increased MAE error. This highlights that reconstruction helps the compressed tokens retain geometric fidelity critical for control alignment. Disabling PE alignment (*W/O_PE*) yields drops in Dynamic Degree and controllability, indicating that aligned spatiotemporal coordinates facilitate smoother motion and consistent object placement across frames. Disabling causal separation (*W/O_Causal*) causes a noticeable latency increase, confirming that causal masking is essential for KV caching and efficient inference. Interestingly, non-causal variants sometimes improve video metrics due to additional bidirectional context, but at the cost of losing the efficiency benefits that motivate ASC-MoT. Removing compression entirely (*W/O_Compress*) restores near-full controllability but increases latency.

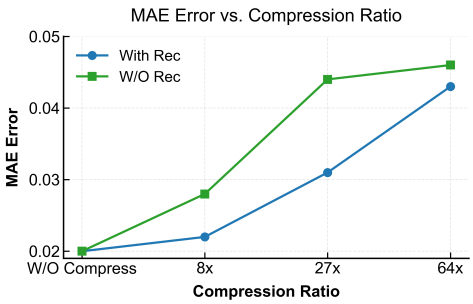

Figure 4: **MAE Error vs. Compression Ratio.** This plot illustrates the trade-off between controllability and token compression.

Ablation on Token Compression Ratio. We visualize the MAE error across various token compression scales to evaluate structural fidelity in Figure 4. While increasing the compression ratio naturally leads to a rise in MAE error, our Reconstructive Condition Compression (RCC) significantly mitigates this degradation compared to the baseline without the reconstruction objective (W/O Rec). Notably, at an $8\times$ compression ratio, the model with RCC maintains a precision level nearly on par with the uncompressed version, demonstrating that our reconstructive bottleneck effectively anchors fine-grained structural information even within highly compact latent sequences.

Interestingly, we also observe a trend regarding compression strength in the structural control task: as the compression ratio increases, controllability decreases (higher MAE error), whereas perceptual video quality metrics (e.g., aesthetic and imaging quality) tend to slightly improve. This suggests that aggressive compression removes condition-specific structural signals while retaining global statistics, allowing the model more freedom to produce visually appealing textures, but at the cost of weakened control alignment.

We further provide a qualitative comparison of different token compression ratios, with and without the reconstruction loss, in Appendix B.

5 Conclusion

In this paper, we introduced **EasyVideoControl**, an efficient framework for conditional modeling in video diffusion transformers. The Asymmetric Shared Causal Mixture-of-Transformers enables effective unidirectional conditioning with one-pass KV caching, adding only 5–8% latency over unconditioned pipelines and running about $2.9\times$ faster than VACE (17.7s vs. 51.0s; 10.5s under the identical 1.3B backbone). The Reconstructive Conditioning Compression module reduces the conditional context length by $8\times$ while preserving control fidelity (MAE 0.022 vs. 0.020 without compression); beyond $8\times$, further compression yields negligible speedups at a clear fidelity cost. Extensive experiments demonstrate that EasyVideoControl achieves a state-of-the-art trade-off between controllability, quality, and efficiency, marking a step towards scalable and practical controllable video generation.

REFERENCES

Yuxuan Bian, Xin Chen, Zenan Li, Tiancheng Zhi, Shen Sang, Linjie Luo, and Qiang Xu. Video-as-prompt: Unified semantic control for video generation. *arXiv preprint arXiv:2510.20888*, 2025a.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. *arXiv preprint arXiv:2503.05639*, 2025b.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

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

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.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023a.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv e-prints*, pp. arXiv–2305, 2023b.

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.

Haopeng Fang, Di Qiu, Binjie Mao, Pengfei Yan, and He Tang. Motioncharacter: Identity-preserving and motion controllable human video generation. *arXiv preprint arXiv:2411.18281*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

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. *arXiv preprint arXiv:2311.17117*, 2023.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6689–6700, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Denis Karachev. Wan2.2 controlnet, 2025. URL https://github.com/TheDenk/wan2.2-controlnet.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Kling AI Team. Kling: Image-to-video special effects. https://klingai.kuaishou.com/image-to-video/special-effects/new, 2025.

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.

Baolu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, et al. Vfxmaster: Unlocking dynamic visual effect generation via in-context learning. *arXiv preprint arXiv:2510.25772*, 2025.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. *Advances in Neural Information Processing Systems*, 37:39869–39900, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Zhengwentai Sun. clip-score: CLIP Score for PyTorch. `https://github.com/taited/clip-score`, March 2023. Version 0.2.1.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*, 2025.

Vidu Team. Vidu: AI video templates. `https://www.vidu.com/templates`, 2025.

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.

Cong Wang, Jiaxi Gu, Panwen Hu, Haoyu Zhao, Yuanfan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. Easycontrol: Transfer controlnet to video diffusion for controllable generation and interpolation. *arXiv preprint arXiv:2408.13005*, 2024.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025a.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023c.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*, 2024.

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing. *arXiv preprint arXiv:2506.04216*, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pp. 1–15, 2024a.

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*, pp. 3836–3847, 2023.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *ICLR*, 2024b.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024c.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19513–19524, 2025a.

Zhiyuan Zhang, Can Wang, Dongdong Chen, and Jing Liao. Flextraj: Image-to-video generation with flexible point trajectory control. *arXiv preprint arXiv:2510.08527*, 2025b.

Appendix

AI USE STATEMENT

In this work, generative AI tools were used only for language polishing of the manuscript, including improving readability and fixing grammatical issues. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

A PRELIMINARIES: VIDEO DIFFUSION TRANSFORMERS

A standard Video Diffusion Transformer operates on a sequence of latent tokens encoded from video frames by a pretrained VAE encoder. The input to the transformer is a sequence of tokens $X = [x_1, x_2, \ldots, x_{N_v}]$, where N_v is the sequence length. To capture complex temporal dynamics, these tokens are processed through a series of DiT blocks comprising self-attention, cross-attention, and feed-forward networks (FFN). Rather than conventional noise prediction, the model is optimized by minimizing the flow matching loss Lipman et al. (2022), which trains the network to predict the velocity field v_t that defines a probability path between the Gaussian noise and the data distribution. Specifically, the model learns a rectified flow by minimizing the objective:

$$\mathcal{L} = \mathbb{E}_{t,z_0,z_1}[\|v_\theta(z_t, t, c) - (z_1 - z_0)\|^2], \tag{8}$$

where $z_t = (1-t)z_0 + tz_1$ denotes the linear interpolation between the data z_0 and noise z_1, and c denotes the conditioning inputs (e.g., text embeddings and control signals). In this framework, the Transformer backbone serves as a powerful prior for spatiotemporal synthesis, into which external conditioning signals can be injected to steer the generation process towards specific structural or semantic requirements.

B ADDITIONAL EXPERIMENTS

We report additional experiments that complement the evaluation in the main paper: (1) a same-backbone comparison with VACE, (2) an average-pooling baseline for the RCC module, and (3) LoRA rank ablations (Table 2); (4) a detailed efficiency analysis in terms of peak memory, latency, and FLOPs (Table 3); and (5) a qualitative ablation on token compression ratios (Figure 5).

Same-backbone comparison with VACE. Since different open-source baselines are built on different backbones, a perfectly unified comparison of all methods under exactly the same backbone, VAE, and inference pipeline is often infeasible in practice. To further address this fairness concern, we additionally train our method on the Wan2.1-1.3B backbone and compare it with VACE (Wan2.1-1.3B) under identical base model conditions. As shown in the first block of Table 2, under the same backbone our method attains video quality and controllability comparable to VACE, while reducing latency from 51.0s to 10.5s and trainable parameters from 100% to 10%.

RCC vs. average pooling. RCC differs from naive pooling in that token aggregation is learned via local spatiotemporal self-attention rather than fixed operations. We validate this distinction with an additional average-pooling baseline ($8\times$ compression, without reconstruction, 5B backbone) in the second block of Table 2. RCC consistently outperforms average pooling on controllability (MAE 0.028 vs. 0.035) and on most video quality and consistency metrics at identical latency.

LoRA rank ablation. The third block of Table 2 reports LoRA rank ablations ($r \in \{64, 128, 256\}$) for depth-controlled T2V at $8\times$ compression on the 5B backbone. Rank 256 achieves the lowest MAE (0.021), while rank 128 attains a nearly identical MAE (0.022) at a lower parameter cost (10% vs. 14% trainable parameters); rank 64 is slightly worse. We therefore adopt rank 128 by default as the balance between parameter efficiency and control precision.

Method	Video Quality & Consistency ↑							Ctrl. ↓	Efficiency ↓	
	Aes.	Bkg.	Dyn.	Img.	Mot.	Ovl.	Tmp.	MAE	Lat.	Par.
Same-backbone (Wan2.1-1.3B, depth T2V)										
VACE (1.3B)	**61.04**	**95.90**	70.00	**67.94**	97.91	25.40	96.31	0.026	51.0s	100%
Ours (1.3B)	59.91	94.79	**70.00**	64.35	**98.58**	**25.42**	**97.47**	**0.023**	**10.5s**	**10%**
RCC vs. average pooling (8×, w/o reconstruction, 5B)										
AvgPool	**59.63**	95.34	**65.00**	60.28	97.13	25.82	96.59	0.035	17.7s	10%
RCC	59.56	**96.37**	60.00	**60.50**	**98.60**	**25.95**	**96.94**	**0.028**	17.7s	10%
LoRA rank ablation (8×, 5B)										
Rank64	**58.84**	94.58	65.00	**60.19**	97.27	**25.83**	95.59	0.030	17.7s	8%
Rank128	58.15	**96.24**	65.00	59.65	**98.67**	25.51	**97.08**	0.022	17.7s	10%
Rank256	57.94	95.80	**70.00**	59.22	98.16	25.32	96.88	**0.021**	17.7s	14%

Table 2: **Additional comparisons and ablations.** (1) Ours (8×) vs. VACE with the same backbone (Wan2.1-1.3B); (2) RCC vs. average pooling (both without reconstruction); (3) LoRA rank ablations. All training and evaluation (VBench) settings follow the main paper for the depth-control T2V task unless otherwise specified. **Bold**: best, underline: second best.

Efficiency analysis. Table 3 provides a comprehensive efficiency analysis comparing peak memory, latency, and total FLOPs. Our full model (8×) achieves the lowest latency and the lowest FLOPs among all compared methods for both T2V and I2V settings. The two key designs are complementary: disabling causal masking (W/O Causal) prevents KV caching and raises latency by 58% (17.7s → 27.9s), while removing compression (W/O Compress) adds 17% latency overhead (17.7s → 20.7s) and increases peak memory.

Method	Resolution	Peak Memory (GB)		Latency (s)		FLOPs	
		T2V	I2V	T2V	I2V	T2V	I2V
ControlNet	480×832×49	25.5	-	38.7	-	4.09e15	-
VACE	480×832×49	**15.6**	-	51.0	-	2.56e16	-
VAP	480×832×49	-	**44.9**	-	317	-	5.85e16
W/O Causal	480×832×49	48.9	50.8	27.9	28.0	3.99e15	4.03e15
W/O Compress	480×832×49	51.7	52.4	20.7	22.4	3.97e15	4.01e15
Ours (8×)	480×832×49	49.2	49.97	**17.7**	**19.4**	**3.42e15**	**3.46e15**

Table 3: **Efficiency analysis.** Memory and latency measured on a single H800 GPU. FLOPs: total FLOPs accounting for 50 denoising steps, matching the sampling protocol used throughout the main paper. **Bold**: best, underline: second best. **Backbones:** ControlNet is a 7% skip-layer variant built on Wan2.2, VACE uses Wan2.1-1.3B, and our variants use Wan2.2-TI2V-5B; the higher peak memory of our variants primarily reflects the larger base model rather than conditioning overhead.

Choice of compression ratio. Beyond 8× compression, conditional tokens are already a small fraction of the total sequence, so backbone computation dominates; further reduction yields marginal latency gains but significant quality degradation, motivating our choice of 8× compression. We defer the discussion of text-guided token selection to Appendix C.

Qualitative ablation on compression ratio. Figure 5 illustrates the qualitative effects of different token compression ratios, with and without the reconstruction loss. At an 8× compression ratio, our method retains fine-grained controllability: the synthesized frames closely follow the depth condition and remain temporally coherent with high visual quality. Removing the reconstruction loss under the same compression ratio leads to noticeable misalignment (e.g., mismatched dog poses), indicating that reconstruction plays a critical role in preserving control signals. At 27× compression, controllability is still preserved when reconstruction is enabled, whereas disabling it again produces misalignment between motion and depth conditions. At 64× compression, both variants fail to maintain controllability, suggesting that at such extreme compression levels, the conditional representation can no longer retain sufficient structural information for reliable guidance.

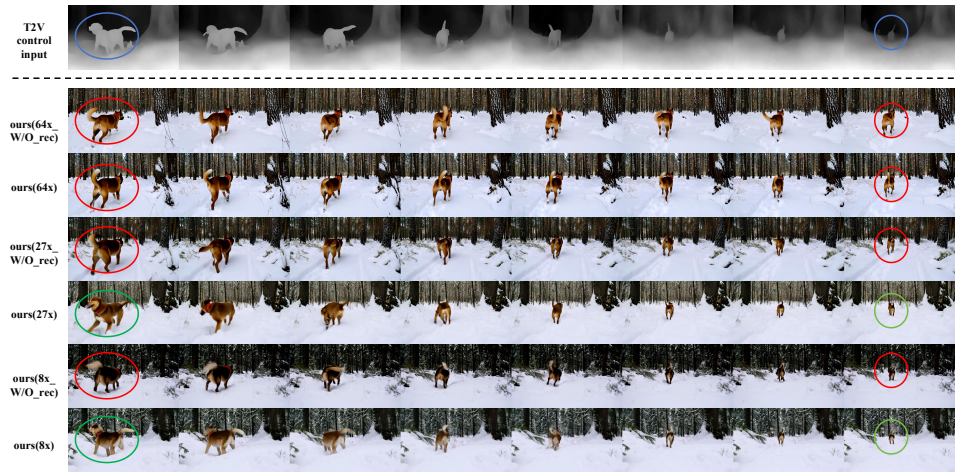

Figure 5: Qualitative comparison of different token compression ratios, with and without the reconstruction loss.

C Limitation and Future Work

Currently, our experiments are primarily conducted on videos with a resolution of 480×832 and a duration of 49 frames. While this configuration covers a wide range of practical applications, longer videos and higher resolutions (e.g., 1080p or 4K) impose more stringent requirements on token compression efficiency and KV cache management. In particular, our full model already consumes 49.2GB (T2V) and 49.97GB (I2V) of peak memory for a single $480\times832\times49$ sample, and moving to 1080p increases the spatial resolution by roughly $5.2\times$, proportionally scaling both the latent token grid and the cached condition KV states across all transformer layers. Practical high-resolution deployment will therefore require dedicated memory management, such as paged KV caching, sliding-window spatiotemporal attention over condition tokens, or sequence parallelism across multiple GPUs. In future work, we plan to explore the scaling laws of our model on ultra-large-scale video datasets and further validate its capability for fine-grained control at higher resolutions. Finally, text-guided token selection—pruning condition tokens irrelevant to the prompt—is a promising direction for further compression; to preserve one-pass KV caching, such selection should be performed offline before sampling (e.g., via an initial text–condition cross-attention pass) rather than as a function of the time-varying latents.

D Additional Results

In Figure 6 and Figure 7 we show more results of Semantic I2V Control. In Figure 8 and Figure 9 we show more results of Structural T2V Control.

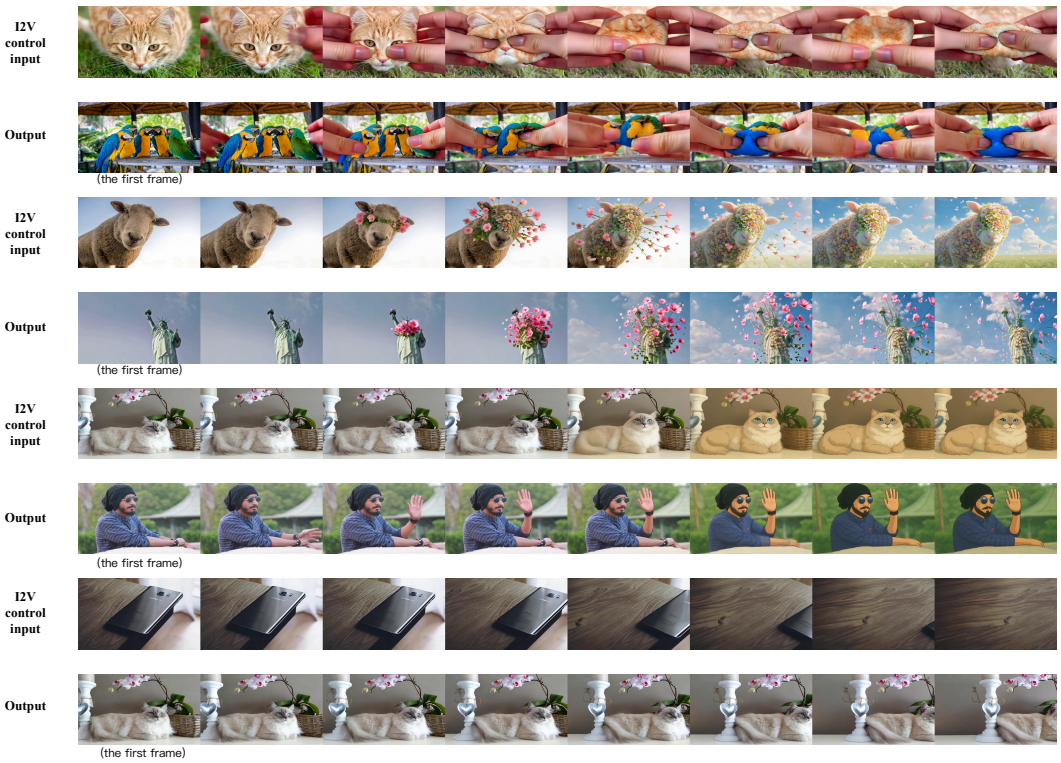

Figure 6: Additional Results of Semantic I2V Control.



Figure 7: Additional Results of Semantic I2V Control.



Figure 8: Additional Results of Structure T2V Control (Openpose).



Figure 9: Additional Results of Structure T2V Control (Scribble).