# Structure From Tracking: Distilling Structure-Preserving Motion for Video Generation

Yang Fei[1] George Stoica[2] Jingyuan Liu[3] Qifeng Chen[1*] Ranjay Krishna[2]
Xiaojuan Wang[2] Benlin Liu[2*]
[1]HKUST [2]University of Washington [3]Adobe

## Abstract

*Reality is a dance between rigid constraints and deformable structures. For video models, that means generating motion that preserves fidelity as well as structure. Despite progress in diffusion models, producing realistic structure-preserving motion remains challenging, especially for articulated and deformable objects such as humans and animals. Scaling training data alone, so far, has failed to resolve physically implausible transitions. Existing approaches rely on conditioning with noisy motion representations, such as optical flow or skeletons extracted using an external imperfect model. To address these challenges, we introduce an algorithm to distill structure-preserving motion priors from an autoregressive video tracking model (SAM2) into a bidirectional video diffusion model (CogVideoX). With our method, we train* `SAM2VideoX`*, which contains two innovations: (1) a bidirectional feature fusion module that extracts global structure-preserving motion priors from a recurrent model like SAM2; (2) a Local Gram Flow loss that aligns how local features move together. Experiments on VBench and in human studies show that* `SAM2VideoX` *delivers consistent gains (+2.60% on VBench, 21–22% lower FVD, and 71.4% human preference) over prior baselines. Specifically, on VBench, we achieve 95.51%, surpassing REPA (92.91%) by 2.60%, and reduce FVD to 360.57, a 21.20% and 22.46% improvement over REPA- and LoRA-finetuning, respectively. The project website can be found at https://sam2videox.github.io/*

## 1. Introduction

From Heraclitus to Bergson, philosophy has cast cognition as the apprehension of *becoming* rather than *being* [1, 9]. While image generation models have excelled at generating what is *there* in high-fidelity images [2, 3, 16], our best video generation still struggles to express the dynamics of *change*. The central challenge is *structure-preserving* motion: dynamics that maintain part topology and local neighborhoods while allowing constrained deformations. Without these constraints, generated motions drift; limbs shear, textures tear, and object identity is lost. Only by achieving this can models move beyond static appearance toward faithful world simulators.

Motion remains a major challenge. This is particularly true for articulated and highly deformable objects, such as humans and animals, where models often suffer from inconsistent or physically implausible transitions in object states. Inference-time interventions, such as ControlNet-style [37] conditioning on explicit motion representations during inference, require knowing the ideal motion apriori. Unlike rigid objects whose motion can be well represented by simple dragged trajectories [7], articulated and deformable entities lack a unified motion representation A common assumption is that low motion quality stems from insufficient training data, especially for high-quality complex articulated motions. However, scaling up or augmenting training data only helps marginally; training with motion proxies (optical flow [4], skeletons [13]) for objects [28, 35] still results in physically implausible transitions. Our experiments show that generated videos still produce lions walking without alternating legs and cyclists with static knees. Scaling up such priors results in noisy training data; motion priors are usually collected using imperfect models (*e.g.* RAFT can generate optical flow priors [29]).

To improve articulated motion, we propose a simple but powerful idea: deriving *structure from tracking*. Our model, `SAM2VideoX`, distills *structure-preserving* motion priors from a video tracking model into a video diffusion generator. Previous work has shown that distilling image representations improves image generation fidelity [38]. To generalize this insight from static to dynamic generation, we distill video representations to improve video generation. Specifically, we leverage SAM2 [23], a state-of-the-art video tracking model trained on large-scale, diverse video data. SAM2 is capable of maintaining object identity across long sequences and through complex occlusions. To track, SAM2's internal representations have captured how parts

*corresponding authors


Figure 1. We present a training algorithm to distill structure-preserving motion priors from SAM2 into a video diffusion model to improve motion accuracy and smoothness in generated videos. Compared to advanced image-to-video models CogVideoX [34] and HunyuanVid [16], our SAM2VideoX produces videos with superior or highly competitive fidelity, despite HunyuanVid having more than twice number of parameters (13B vs. our 5B).

move together, how limbs stay connected, and how occlusions resolve over time. Instead of conditioning generation on explicit control signals like optical flow [4] or skeletons [13], SAM2VideoX extracts implicit structural cues directly from SAM2's internal representations. By transferring these motion priors, the generator acquires an internal sense of structure and continuity. In short, we leverage the structural understanding of a tracker to guide the generation of motion.

However, transferring useful information from SAM2 into a video generation model is technically challenging. We find that directly supervising diffusion models to predict SAM2 output masks yields only limited benefit, as the masks are discrete and boundary-focused; they fail to supervise useful fine-grained motion. Also, a direct alignment between feature spaces is hindered by an architectural asymmetry: state-of-the-art video generation models use DiT [20] architecture with bidirectional attention to access global context, while SAM2 is inherently recurrent and causal. To bridge this gap, we extract a supervision signal by fusing *forward* and *backward* SAM2 features, where backward features are extracted by reversing the order of frames in a training video. Together, the forward and backward features represent a better global video context. We align video diffusion features with this supervisory signal using a *Local Gram Flow* loss. Although $\ell_2$ loss has worked well for image generation [38], we find that a local Gram loss captures better motion priors by emphasizing local relational structure.

Across qualitative and quantitative evaluations, SAM2VideoX yields more realistic, structurally coherent motion. On VBench [12] (matched dynamic degree) we achieve 95.51%, surpassing REPA [36] (92.91%) by 2.60 points. Our FVD is 360.57, a reduction of 21.20% and 22.46% versus REPA- and LoRA-finetuning [11]. Since existing benchmarks are limited in assessing preservation of structure in articulated motion, we further conduct a human evaluation, where 71.4% of ratings prefer our results. Qualitatively, SAM2VideoX produces videos with the correct number of legs when animals walk, ensuring plausible human limb trajectories during complex activities, and producing accurate human-object interactions. These gains indicate that SAM2-guided distillation substantially strengthens structure-preserving motion in video diffusion models without sacrificing visual quality.

## 2. Related Work

**Video diffusion models.** Video diffusion models have progressed rapidly, spanning both UNet-based [24] architectures such as Stable Video Diffusion [2], and DiT-based [20] models including Sora [3], CogVideoX [34], HunyuanVid [16], OpenSora [21], Open-Sora-Plan [17], Wan-Video [32], Cosmos [19]. While these models can generate visually impressive videos, producing realistic and coherent structure-preserving motion remains challenging, especially for articulated and deformable entities. Given the difficulty of designing reliable inference-time control signals [37] for such objects, our goal is to enhance the base model's intrinsic ability to generate structure-preserving motion without *relying on auxiliary handcrafted motion controls during inference*.

**Motion understanding.** Understanding motion has long been central to video analysis [10, 22]. Point trajectories and optical flow [14, 29] describe local, adjacent-frame changes and carry limited semantics; they often degrade un-

der fast or long-range motion and struggle to maintain object identity through occlusions. Mask-based tracking offers instance-level signals that are more stable in cluttered scenes. SAM2 [23] tracks user-prompted regions across long sequences and is known to generalize well across domains while preserving object identity through occlusions. However, raw masks are discrete and boundary-focused, which discards much of the appearance and motion structure that video diffusion models need. We therefore leverage SAM2's internal features as motion priors: they are dense, continuous and temporally consistent, and they provide long-range correspondences that are useful for structure preservation.

**Representation alignment.** Representation alignment was introduced for image generation by REPA [36] and inspired several follow-ups. In the video domain, two main approaches exist: aligning diffusion features to structured motion signals such as trajectories or flow [4, 13]; or to generic video encoders like VideoMAEv2 [30], as in VideoREPA [38]. Both suffer from key limitations. First, Trajectories and flow provide local supervision and are sensitive to long-range dynamics, which weakens their usefulness for structure preservation. Second, popular video encoders like VideoMAEv2 are optimized for high-level semantic tasks rather than low-level motion understanding. In contrast, our approach aligns diffusion features to SAM2's internal features. This prior is unified and structure-centric, allowing transfer across humans and animals while yielding object-consistent motion representations without requiring a specific controller at inference.

## 3. Preliminaries

In this work, we employ diffusion transformer (DiT) [20] as our base video generation model, aiming to distill SAM2's motion understanding into the DiT to enhance its ability to learn structure-preserving motion from in-the-wild videos. We first introduce preliminary background on latent video diffusion models and the SAM2 architecture.

**Latent Video Diffusion Model.** Diffusion models generate samples by inverting a forward noising process [8, 26, 27]. In the latent setting, a pre-trained autoencoder, with encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$, maps a video $\mathbf{x} = \{I_0, \dots, I_{N-1}\} \in \mathbb{R}^{N\times H\times W\times C}$ to a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$, where $z \in \mathbb{R}^{N'\times H'\times W'\times C'}$. At timestep $t$, noisy latent $\mathbf{z}_t$ is sampled as

$$\mathbf{z}_t = \alpha_t\,\mathbf{z} + \sigma_t\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

A DiT $f_\theta$ is trained to predict the velocity target $\mathbf{v} = \alpha_t\boldsymbol{\epsilon} - \sigma_t\mathbf{z}$ under the standard v-prediction objective. At inference, iterative denoising maps $\mathbf{z}_T \to \mathbf{z}_0$, after which the decoder $\mathcal{D}(\mathbf{z}_0)$ produces the final video.

**Segment-Anything Model 2 (SAM2).** The Segment-Anything Model 2 (SAM2) extends the image segmentation capability of SAM 1 [15] to the video domain, i.e., tracking an object in a video conditioned on given prompts such as points or boxes.

Given a video $\mathbf{x} = \{I_0, I_1, ..., I_{N-1}\}$, SAM2 processes the video recurrently and produces a segmentation mask for each frame. Specifically, an image encoder $\mathcal{I}$ extracts frame embedding $F$ from the current frame, which is then enhanced by a memory attention module $\mathcal{M}$ aggregating historical context from a memory bank $\mathcal{B}$ to produce $F_{\text{mem}}$. A mask decoder $\mathcal{D}_{\text{mask}}$ subsequently takes $F_{\text{mem}}$ and user prompts to generate the segmentation mask. By applying this recurrent procedure across all frames, SAM2 produces both a sequence of masks:

$$\mathbf{M} = \{M_0, M_1, ..., M_{N-1}\}$$

and a sequence of memory features from $\mathcal{M}$:

$$\mathbf{F}_{\text{mem}} = \{F_{\text{mem},0}, F_{\text{mem},1}, ..., F_{\text{mem},N-1}\}.$$

## 4. Method

We distill SAM2's motion structure prior into the video DiT model by aligning their internal feature representations. This is achieved through a learnable feature alignment module (Sec. 4.1) that projects intermediate DiT features into a latent space where they can be matched to those of SAM2. To align their relational motion structures, we introduce a novel Local Gram Flow (LGF) feature matching operator (Sec. 4.2). Furthermore, due to the autoregressive nature of SAM2'feature, in contrast to the bidirectional features in DiT, we propose a method that combines SAM2's forward and backward video features into a single bidirectional representation, providing a more suitable teacher signal to be distilled from (Sec. 4.3). Finally, a Local Gram Flow motion distillation loss (Sec. 4.4) is applied to enforce this alignment in the latent space. Fig. 2 provides an overview of our full method.

### 4.1. Feature Alignment Network

Formally, given a training video $\mathbf{x}$, we encode it to a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$ using the video VAE, then add noise at timestep $t$ to produce $\mathbf{z}_t$. The noised latent $\mathbf{z}_t$ and timestep $t$ are then fed into the denoising network $f_\theta$, from which we extract intermediate activations as video diffusion features $\mathbf{F}_{\text{diff}} \in \mathbb{R}^{N'\times H'\times W'\times C'}$, where $N'$ is the number of latent frames. Here $\mathbf{F}_{\text{diff}}$ denotes the activations from a selected intermediate layer. For the same video, we extract SAM2's internal features $\mathbf{F}_{\text{SAM2}}$ as a distillation teacher. Compared with the output segmentation masks, the internal feature representations provide richer spatio-temporal information. They capture object motion and part-level dynamics and can teach diffusion model internalize motion priors beyond simple boundary cues.

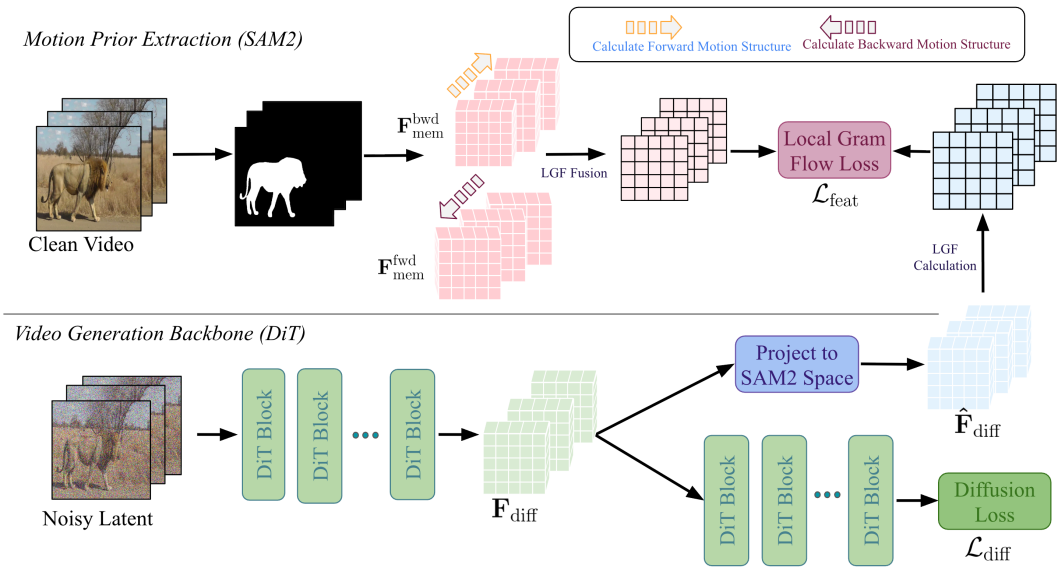


Figure 2. **Method overview.** The framework consists of two parallel branches: **(Top)** The *Motion Prior Extraction* branch extracts forward and backward memory features ($\mathbf{F}^{\text{fwd}}_{\text{mem}}, \mathbf{F}^{\text{bwd}}_{\text{mem}}$) from SAM2 given a clean video, and fuses them into a bidirectional teacher representation. **(Bottom)** The *Video Generation Backbone* takes noisy latents as input, and the intermediate DiT features $\mathbf{F}_{\text{diff}}$ are projected into the SAM2 space as $\hat{\mathbf{F}}_{\text{diff}}$. Then the proposed **Local Gram Flow loss** ($\mathcal{L}_{\text{feat}}$) is used to align the spatio-temporal structure of the projected student features with the teacher priors.

To align $\mathbf{F}_{\text{diff}}$ and $\mathbf{F}_{\text{SAM2}}$, We project $\mathbf{F}_{\text{diff}}$ into SAM2's feature space. Specifically, we add a projection module $\mathcal{P}$ on top of $\mathbf{F}_{\text{diff}}$, which consists of an interpolation layer with skip connections for temporal dimension matching, and then followed by a three-layer MLP, yielding:

$$\hat{\mathbf{F}}_{\text{diff}} = \mathcal{P}(\mathbf{F}_{\text{diff}}).$$

We then compute a motion distillation loss between $\hat{\mathbf{F}}_{\text{diff}}$ and $\mathbf{F}_{\text{SAM2}}$. Our final objective combines the alignment loss with the standard diffusion loss:

$$\min_{f_\theta, \mathcal{P}} \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{feat}}(\hat{\mathbf{F}}_{\text{diff}}, \mathbf{F}_{\text{SAM2}}),$$

where $\mathcal{L}_{\text{diff}}$ is the v-prediction loss and $\lambda = 0.5$ balances the terms.

## 4.2. Local Gram Flow Feature Matching

To capture cross-frame spatio-temporal motion structure, rather than performing direct one-to-one feature matching between $\hat{\mathbf{F}}_{\text{diff}}$ and $\mathbf{F}_{\text{SAM2}}$, we instead match their respective *Gram matrices*, which encode pairwise dot products of token feature embeddings, *i.e.*, pairwise token similarities. However, computing the full Gram matrices is computationally prohibitive for video diffusion models due to the large number of tokens. We therefore propose *Local Gram Flow*, which computes the dot products only between each token and the tokens within its $7 \times 7$ spatial neighborhood in the subsequent frame (see Fig. 3). This yields local similarity vectors at each position that model likely motion trajectories to the next frame. We denote this operator as $\text{LGF}(\cdot)$.

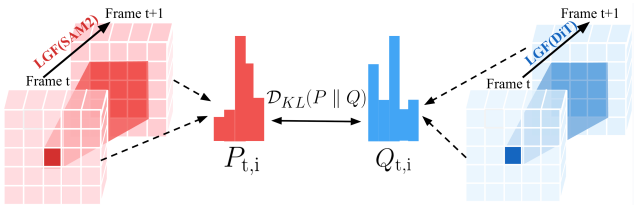


Figure 3. **Illustration of the Local Gram Flow (LGF) Loss.** The LGF operator captures motion structure by computing similarities between a token at frame $t$ and its 7x7 spatial neighborhood in the subsequent frame $t+1$. Instead of matching absolute values (e.g., $\ell_2$), we convert the resulting similarity vectors ($P_{\text{t,i}}, Q_{\text{t,i}}$) into probability distributions and align them using the KL Divergence. This forces the model to learn relative motion patterns, not just feature values.

## 4.3. Bidirectional Fusion of Causal SAM2 Features

A key challenge in aligning the feature representation of DiT and SAM2 is the architectural asymmetry between them. Video DiTs typically employ bidirectional attention that allows each token to attend to all frames, whereas SAM2's recurrent mechanism constrains its memory feature $\mathbf{F}_{\text{mem}}$ at each timestep to encode only current and past frames. To bridge this gap and create a teacher feature $\mathbf{F}_{\text{SAM2}}$ that is aware of the full video context (similar to

DiT), we construct it from both a forward and backward pass of SAM2. The backward pass is obtained by feeding the temporally reversed video into SAM2 to extract its backward features. After obtaining the backward features, we remap them to the original temporal order (i.e., $t \mapsto N-1-t$) to align with the forward features, producing:

$$\mathbf{F}_{\text{mem}}^{\text{fwd}} = \{F_{\text{mem},0}^{\text{fwd}}, F_{\text{mem},1}^{\text{fwd}}, \ldots, F_{\text{mem},N-1}^{\text{fwd}}\},$$

$$\mathbf{F}_{\text{mem}}^{\text{bwd}} = \{F_{\text{mem},0}^{\text{bwd}}, F_{\text{mem},1}^{\text{bwd}}, \ldots, F_{\text{mem},N-1}^{\text{bwd}}\}.$$

Empirically, using separate projectors to align separately to $\mathbf{F}_{\text{mem}}^{\text{fwd}}$ and $\mathbf{F}_{\text{mem}}^{\text{bwd}}$ provides marginal improvement, as gradient conflicts destabilize training. We therefore fuse them into a unified bidirectional teacher feature.

However, fusing these two features is non-trivial. As our ablation study demonstrate (Table 2), naively adding them ($k\mathbf{F}_{\text{mem}}^{\text{fwd}} + (1-k)\mathbf{F}_{\text{mem}}^{\text{bwd}}$) leads to severe performance degradation. Because the final fused feature will be aligned with the projected DiT feature through their Local Gram Flows, we instead directly fuse their Local Gram Flows via a convex combination, which stabilizes training while preserving complementary information:

$$\text{LGF}(\mathbf{F}_{\text{SAM2}}) = \text{k LGF}(\mathbf{F}_{\text{mem}}^{\text{fwd}}) + (1 - \text{k}) \text{ LGF}(\mathbf{F}_{\text{mem}}^{\text{bwd}})$$

### 4.4. Motion Distillation Loss

Finally, the motion distillation loss, $\mathcal{L}_{\text{feat}}$, is designed to match the LGF distributions of the student $\hat{\mathbf{F}}_{\text{diff}}$ and the fused teacher $\mathbf{F}_{\text{SAM2}}$, rather than enforcing one-to-one correspondence as in a standard $\ell 2$ loss. More specifically, we align the distribution of spatio-temporal motion similarities between the video DiT feature and the SAM2 feature. We therefore apply a softmax to each token's similarity vector (turning it into a probability distribution) and measure the distance using the KL divergence. This approach focuses on the relative ranking of similarities, which we argue better captures the underlying motion structure. Specifically, we compute probability distributions:

$$\mathbf{P} = \text{S}\Big(\text{LGF}(\mathbf{F}_{\text{SAM2}})\Big),$$

$$\mathbf{Q} = \text{S}\Big(\text{LGF}(\hat{\mathbf{F}}_{\text{diff}})\Big).$$

The motion distillation loss averages the KL divergence over all spatial tokens $\Omega$ and all $N' - 1$ latent frames for which LGF is computed:

$$\mathcal{L}_{\text{feat}}\Big(\hat{\mathbf{F}}_{\text{diff}}, \mathbf{F}_{\text{SAM2}}\Big) = \frac{1}{(N'-1)|\Omega|} \sum_{t=0}^{N'-2} \sum_{i \in \Omega} \text{KL}\big(P_{\text{t,i}} \,\big\|\, Q_{\text{t,i}}\big)$$

Our ablation studies (Table 2) empirically validate that this LGF-KL combination is critical, yielding significant gains over simpler alternatives (e.g., LGF with $\ell 2$ loss, or direct feature matching).

## 5. Experiments

**Dataset & training configuration.** We curate a motion-focused dataset of 9,837 single-subject video clips from open-source video generation datasets(Panda70M [5], MM-Trailer [6], MotionVid [33]), capturing diverse motion patterns across animals and humans. All videos are at 8 FPS, capped at 100 frames. Our approach builds upon CogVideoX-5B-I2V [34], extracting intermediate DiT features from the 25th block output as $F_{\text{diff}}$. We first obtain object bounding boxes using GroundingDINO [18] to prompt SAM2 mask generation. To align with DiT features, we propagate subject masks from both temporal directions: using the first-frame mask to compute forward features $\mathbf{F}_{\text{mem}}^{\text{fwd}}$ and the last-frame mask for backward features $\mathbf{F}_{\text{mem}}^{\text{bwd}}$. To avoid SAM2 inference overhead during training, we precompute features for clips starting at every 20 frames.

We implement LoRA fine-tuning with rank 256 and scaling factor $\alpha = 128$. Training uses AdamW optimization with learning rate $1 \times 10^{-4}$ and momentum parameter $(\beta_1, \beta_2) = (0.9, 0.95)$. We train for 3,000 steps on $8\times$H200 GPUs, with global batch size 32 through gradient accumulation over four steps per GPU.

**Baselines.** We compare our complete approach, which uses bidirectional feature fusion and Local Gram Flow loss, with four baselines: (1) CogVideoX-5B-I2V as the base model; (2) "+ LoRA fine-tuning", which simply fine-tunes the base model with LoRA on our curated motion dataset; (3) "+ Mask supervision", which adds a linear projection layer on top of our projection head $\mathcal{P}$ to directly predict the subject mask, trained with mask loss supervision instead of feature alignment; (4) "+REPA" , which utilizes REPA Loss [36] to align DiT features with external DINOv3 [25] features.

**Evaluation protocol.** We evaluate across three complementary axes: objective motion metrics, perceptual quality, and human preference. For objective evaluation, we filter 85 images and compute four metrics from the VBench-I2V suite [12]: Motion Smoothness, Subject Consistency, and Background Consistency, and Dynamic Degree. We define consistency metrics as measures of temporal structure stability, while Dynamic Degree quantifies the magnitude of motion. Since consistency scores often correlate negatively with motion magnitude(i.e., static videos trivially achieve perfect consistency), to ensure a fair comparison of motion quality, we exclude baselines with a lower Dynamic Degree than the base model. The overall Motion Score is obtained by averaging the smoothness and consistency metrics (min-max normalized). For the Extended Motion Score, we incorporate I2V-Subject and I2V-Background Consistency with a 0.5 weight, following the official VBench protocol. For perceptual quality, we compute Fréchet Video Distance (FVD) [31] on a separate set of 200 videos randomly sampled from the training dataset. For human preference, we

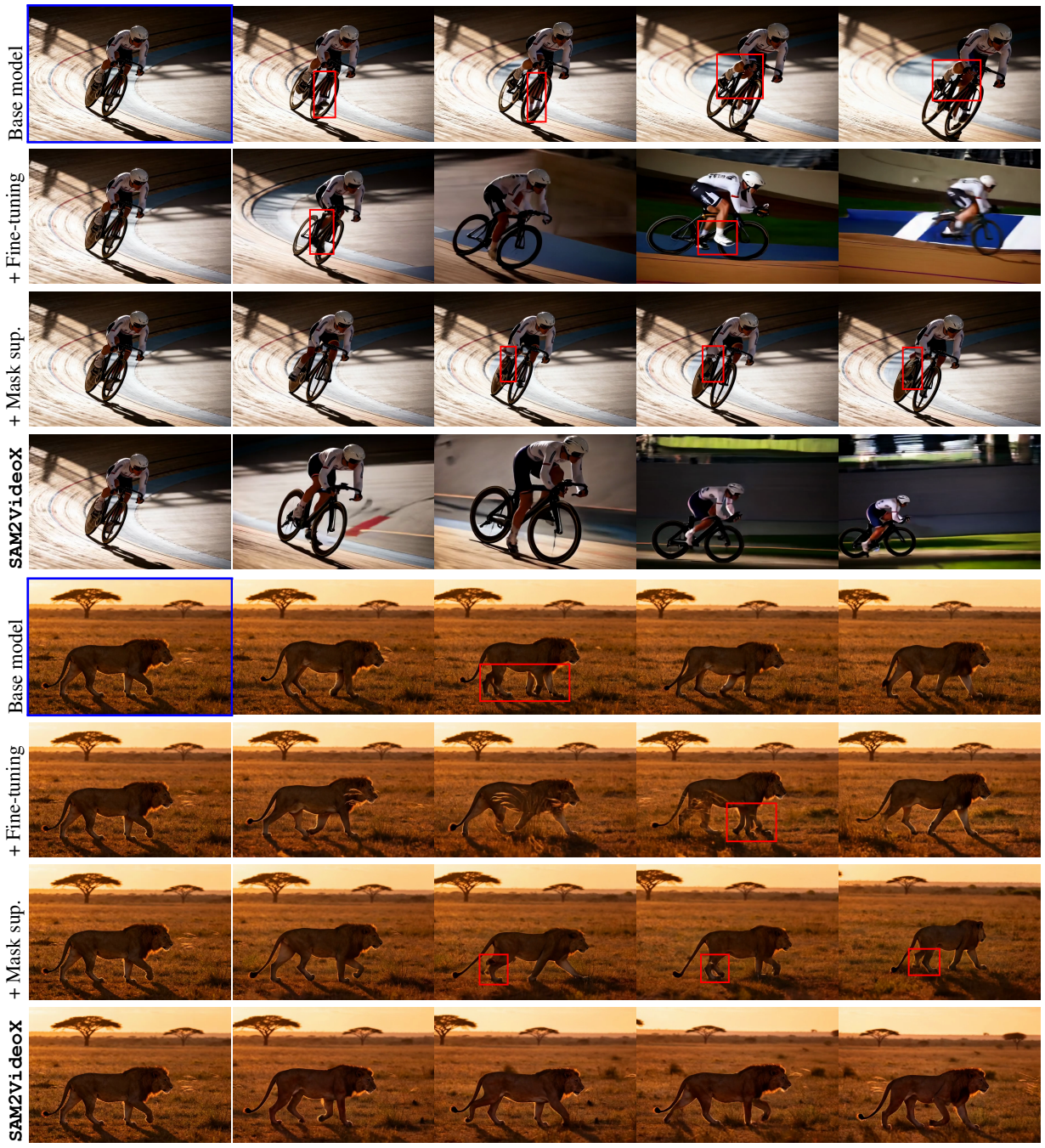


Figure 4. **Qualitative comparison on articulated motion.** The blue box indicates the input image. Red boxes highlight common failure modes in baselines, including structural distortion (cyclist's legs) and physical implausibility (three-legged lion, inconsistent leg motion). In contrast, our method (SAM2VideoX) consistently maintains structural integrity. "+ Fine-tuning" is the LoRA fine-tuning baseline; "+ Mask sup." is the mask supervision baseline. We recommend viewing the supplementary videos.

conduct a double-blind user study using 40 randomly sampled prompts. Participants are presented with two side-by-side videos (Ours vs. Baseline) in random order and asked to select the preferred one based on motion smoothness and subject consistency. All methods generate 49-frame videos at 8 fps and 720×480 resolution, using 50 denoising steps with guidance scale 6.0.

Table 1. **Quantitative comparison on established video generation benchmarks.** Our method outperforms all fine-tuning baselines and achieves performance comparable to the strong open-source model HunyuanVid, while significantly surpassing it in perceptual quality (FVD). **BC**: Background Consistency; **SC**: Subject Consistency; **MS**: Motion Smoothness. Extended Motion Score incorporates I2V Subject and Background Consistency (weight 0.5). Higher Motion/Extended Motion Scores indicate better structure preservation; lower FVD indicates superior perceptual quality. Baselines with Dynamic Degree lower than the base model are excluded from VBench comparisons to ensure fairness.

| Method | BC | SC | MS | Motion Score↑ | Ext Motion Score↑ | FVD↓ |
|---|---|---|---|---|---|---|
| CogVideoX (base model) | 97.30 | 94.43 | 98.17 | 94.80 | 95.50 | 660.29 |
| + LoRA Fine-tuning | 97.44 | 93.47 | 97.76 | 94.02 | 94.74 | 465.00 |
| + Mask Supervision | - | - | - | - | - | 397.73 |
| + REPA | 97.41 | 91.99 | 97.31 | 92.91 | 93.77 | 457.59 |
| + **SAM2VideoX (Ours)** | **97.88** | **94.76** | **98.45** | **95.51** | **96.03** | **360.57** |
| HunyuanVid | 96.85 | 95.32 | 98.76 | 95.62 | 96.24 | 583.99 |

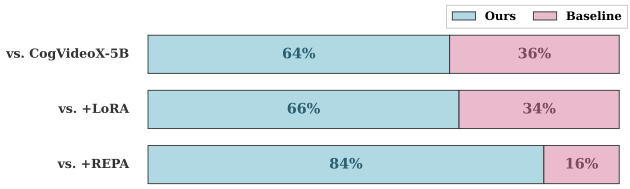


Figure 5. **Human preference win rates.** Participants were shown pairwise comparisons and asked to select the video with 'superior limb consistency and fewer motion artifacts'. Our method was strongly preferred ($> 64\%$) over all baselines, confirming its superior ability to generate plausible, artifact-free motion.

### 5.1. Results

**SAM2VideoX achieves superior structure preservation and perceptual quality compared to all baselines.** As shown in Table 1, our method outperforms the base CogVideoX model and other fine-tuning strategies across all reported metrics. Notably, we achieve an FVD of 360.57, a substantial reduction compared to the strongest baseline (LoRA Fine-tuning at 465.00) and the REPA baseline (457.59). This indicates that our model generates videos with significantly higher perceptual fidelity. Furthermore, our method attains the highest scores in Motion Score (95.51) and Extended Motion Score (96.03), confirming that distilling internal priors from SAM2 effectively suppresses the temporal flickering and identity degradation often observed in standard video diffusion models.

**Feature distillation outperforms coarse mask supervision.** We compare our feature-level alignment against a baseline trained with explicit mask supervision ("+ Mask Supervision"). We build this baseline by adding a linear projection layer on top of $\mathcal{P}$ to predict the mask. While mask supervision delineates subject boundaries, it lacks the fine-grained internal correspondence information necessary for articulating complex motion. Consequently, the mask supervision baseline suffers from severe structural artifacts (e.g., cyclist and lion legs in Fig. 4) and achieves a poorer FVD score of 397.73 compared to our 360.57. Note that we exclude the mask supervision baseline from VBench consistency metrics in Table 1 because it collapses towards static generation (Dynamic Degree 44.59 vs. Base model 45.95), which would yield artificially inflated consistency scores.

**Video-aware priors are essential for temporal consistency.** To validate the necessity of using a video foundation model (SAM2) as the teacher, we compare against REPA [36], which aligns DiT features with the image-based DINOv3 encoder. As DINO is trained on static images, it lacks inherent knowledge of temporal continuity. This limitation is reflected in Table 1, where REPA yields a significantly lower Motion Score (92.91) compared to our method (95.51). This result confirms that aligning with SAM2's memory-based features transfers crucial temporal coherence signals that image-only encoders cannot provide.

**Human evaluators consistently prefer our method.** As illustrated in Figure 5, our method achieves the highest win rates in a double-blind user study, outperforming the base model and fine-tuning baselines by a wide margin. Qualitative results in Figure 4 corroborate this preference: while baselines often struggle with limb consistency (e.g., the disappearing/reappearing legs of the cyclist), our method maintains subject identity and structural integrity throughout the sequence. This demonstrates that our bidirectional alignment strategy successfully translates the robust segmentation priors of SAM2 into high-fidelity video generation.

### 5.2. Ablations

We conduct systematic ablation studies on the VBench-I2V 85-image subset to isolate the contribution of our two core components: the LGF fused bidirectional SAM2 teacher and the LGF-KL distillation loss.

| Configurations | Motion Score↑ | Ext Motion Score↑ |
|---|---|---|
| + LoRA only | 94.02 | 94.74 |
| w/o LGF | 94.58 | 95.24 |
| w/o KL | 94.51 | 95.16 |
| w/ Forward-Only Teacher | 95.07 | 95.58 |
| w/ Separate Projectors | 94.68 | 95.24 |
| Feature-Space Fusion | 94.16 | 94.83 |
| **SAM2VideoX (LGF Fusion)** | **95.51** | **96.03** |

Table 2. **Ablation study on core components (VBench-I2V, ↑).** Using a simple $\ell_2$ loss ("w/o LGF") and an $\ell_2$ loss within LGF space ("w/o KL") both yield marginal improvements only, confirming our full LGF-KL's superiority in capturing motion structure. Removing bidirectional processing ("w/ Forward-Only Teacher") or fusion mechanisms ("w/ Separate Projectors") degrades performance, while feature-level adding ("Feature-Space Fusion") underperforms ours ("LGF Fusion"), validating each design choice.

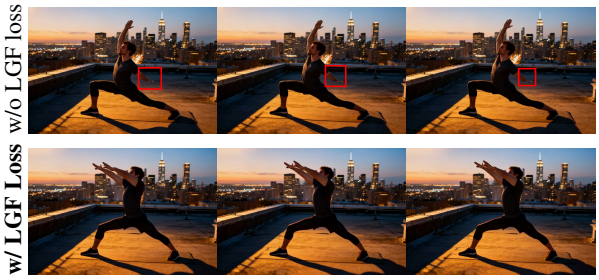


Figure 6. **Qualitative ablation of our LGF-KL Loss.** (Top) Using a standard $\ell_2$ loss ('w/o LGF Loss') on raw features results in visible temporal jitter (note the flickering in the arm). (Bottom) Our full method ('w/ LGF Loss') produces a visibly smoother and more stable motion, validating the design of aligning relational distributions (LGF-KL) over absolute values ($\ell_2$).

**LGF fusion is essential to resolve bidirectional conflicts.** We validate our teacher design in Table 2. While using only the forward stream ($\mathbf{F}^{\text{fwd}}_{\text{mem}}$) yields a strong baseline (Motion Score 95.07), naively incorporating the backward stream via separate projectors causes gradient conflicts, destabilizing training and degrading performance. More critically, fusing forward and backward streams directly in the *feature space* leads to catastrophic collapse (94.16), barely outperforming the LoRA-only baseline. This suggests that raw features from opposite temporal directions interfere destructively. By contrast, our proposed **LGF Fusion** acts as a harmonic integration, resolving these conflicts to achieve the highest score (95.51). Qualitative results in Figure 7 confirm that this relational fusion is key to leveraging bidirectional priors without introducing artifacts.

**KL divergence outperforms $\ell_2$ for structural alignment.** We further ablate the loss function design given our LGF teacher. As shown in Table 2, applying a standard $\ell_2$ loss


Figure 7. **Qualitative ablation of our bidirectional LGF fusion.**(Top) The forward-only teacher ("w/ Fwd-Only") produces artifacts, like the ballerina's arm 'folding' incorrectly (red box). (Middle) "Feat-Space Fusion" causes structural tearing. (Bottom) Our "LGF Fusion" correctly preserves the limb's topological structure, highlighting the necessity of both bidirectional information and a robust fusion strategy.

directly on raw features performs poorly (94.58), proving that strict element-wise alignment is too rigid for transferring high-level motion priors. More revealingly, applying an $\ell_2$ loss within the LGF space performs even worse (94.51). This demonstrates that the LGF operator alone is insufficient; a naive value-based alignment of its relational features fails to capture the correct motion priors. It is the combination of LGF (to capture relational structure) and the KL divergence (to align probabilistic distributions) that is essential. Our full LGF-KL loss achieves the best performance (95.51), confirming that aligning the relative spatio-temporal distributions via KL divergence is superior to forcing exact value matches. Figure 6 visually demonstrates the smoother temporal transitions achieved by this design.

## 6. Conclusion

In this paper, we propose a novel framework that effectively distills the rich, structure-preserving motion priors from SAM2 into video diffusion models. Departing from methods relying on external control signals or limited datasets, we demonstrate that aligning generative features with dense correspondence representations offers a more intrinsic solution to articulated motion generation. Our core contributions—bidirectional feature fusion and the Local Gram Flow loss—enable the seamless transfer of fine-grained motion knowledge without requiring architectural modifications. Extensive experiments validate that our approach not only achieves superior performance on standard benchmarks but also paves the way for leveraging discriminative vision foundation models to enhance generative video dynamics.

## References


[1] Henri Bergson. *Creative Evolution*. Henry Holt and Company, New York, 1911. 1

[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. 1, 2

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 2

[4] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025. 1, 2, 3

[5] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5

[6] Xiaowei Chi, Yatian Wang, Aosong Cheng, Pengjun Fang, Zeyue Tian, Yingqing He, Zhaoyang Liu, Xingqun Qi, Jiahao Pan, Rongyu Zhang, Mengfei Li, Ruibin Yuan, Yanbing Jiang, Wei Xue, Wenhan Luo, Qifeng Chen, Shanghang Zhang, Qifeng Liu, and Yike Guo. Mmtrail: A multimodal trailer video dataset with language and music descriptions, 2024. 5

[7] Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. Dragvideo: Interactive drag-style video editing. In *European Conference on Computer Vision*, pages 183–199. Springer, 2024. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[9] Heraclitus. *Fragments*. 500BC. Translated in *The Presocratic Philosophers*, eds. G. S. Kirk and J. E. Raven, Cambridge University Press, 1957. 1

[10] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[11] 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 *International Conference on Learning Representations*, 2022. 2, 11

[12] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 5, 11

[13] Hyeonho Jeong, Chun-Hao Paul Huang, Jong Chul Ye, Niloy Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. *arXiv preprint arXiv:2412.06016*, 2024. 1, 2, 3, 11

[14] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[16] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1, 2, 14

[17] 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. 2

[18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 5

[19] NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical ai, 2025. 2

[20] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF inter-*

*national conference on computer vision*, pages 4195–4205, 2023. 2, 3

[21] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025. 2

[22] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 2

[23] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 3, 11

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[25] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 5

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[27] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[28] Keqiang Sun, Dor Litvak, Yunzhi Zhang, Hongsheng Li, Jiajun Wu, and Shangzhe Wu. Ponymation: Learning articulated 3d animal motions from unlabeled online videos. In *European Conference on Computer Vision*, pages 100–119. Springer, 2024. 1

[29] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1, 2, 11

[30] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3

[31] 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

[32] 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. 2

[33] Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, et al. Humandreamer: Generating controllable human-motion videos via decoupled generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12391–12401, 2025. 5

[34] 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. 2, 5, 11

[35] Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and Yalin Wang. Omnimotiongpt: animal motion generation with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1249–1259, 2024. 1

[36] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 2, 3, 5, 7

[37] 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. 1, 2

[38] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025. 1, 2, 3

| Method | BC↑ | SC↑ | MS↑ | Motion Score↑ | Ext Motion Score↑ |
|---|---|---|---|---|---|
| Track4Gen* | 97.32 | 94.67 | 98.35 | 95.11 | 95.69 |
| **SAM2VideoX (Ours)** | **97.88** | **94.76** | **98.45** | **95.51** | **96.03** |

Table 3. **Quantitative Analysis of Additional Baseline.** Comparison between point-based control and our feature-based approach. Track4Gen* denotes the modified version adapted to the CogVideoX architecture. Abbreviations: BC (Background Consistency), SC (Subject Consistency), MS (Motion Smoothness).

# A. Implementation Details

**Architecture.** Within the projection head of the Video Feature Alignment module, we employ SiLU activation and Group Normalization across all convolutional and MLP layers. For the interpolation layer, we utilize a temporal interpolation factor of 4, utilized alongside a skip connection with a kernel size of $(3, 1, 1)$ and 768 channels. Subsequently, the widths of the following MLP layers adhere to the sequence $768 \rightarrow 512 \rightarrow 256 \rightarrow 256$.

**Training Details.** We apply LoRA [11] exclusively to the attention modules, keeping all other backbone parameters frozen. For the LoRA parameters, we employ a linear warmup of 200 steps, gradually increasing the learning rate to a peak of $10^{-4}$. For the projector, we utilize a cosine decay scheduler with a 150-step warmup; the learning rate is initialized at $5 \times 10^{-4}$ and decays to a minimum of $1 \times 10^{-5}$. Additionally, we implement a gradient clipping threshold of 1.0.

## A.1. Evaluation Protocol and Metrics

**Evaluation Subset Selection.** To ensure our evaluation aligns with the distribution of our training dataset (which focuses on articulated motion), we curated a subset of 85 prompts from the VBench-I2V [12] benchmark. The selected prompts predominantly feature human or animal subjects. We excluded generic scenery or abstract textures lacking distinct structural subjects, as these samples do not adequately challenge the model's structure-preserving capabilities.

**Metric Selection.** Since our primary objective is structure-preserving video generation rather than artistic creation, we exclude general visual quality metrics such as *Aesthetic Quality* and *Imaging Quality*, as they are less relevant to evaluating structural fidelity. To provide a unified quantitative evaluation, we formulate two composite scores: Motion Score and Extended Motion Score. Following the standard VBench-I2V protocol, we first apply min-max normalization to all individual sub-metrics to map them into a unified range. Let $\hat{s}_{\text{bg}}$, $\hat{s}_{\text{smooth}}$, and $\hat{s}_{\text{subj}}$ denote the normalized scores for Background Consistency, Motion Smoothness, and Subject Consistency, respectively. The Motion Score is defined as the arithmetic mean of these three core structural metrics:

$$S_{\text{motion}} = \frac{\hat{s}_{\text{bg}} + \hat{s}_{\text{smooth}} + \hat{s}_{\text{subj}}}{3} \tag{1}$$

To further account for fidelity to the conditioning image, the Extended Motion Score incorporates I2V consistency metrics. We assign a weight of 0.5 to both I2V Subject Consistency ($\hat{s}_{\text{i2v-s}}$) and I2V Background Consistency ($\hat{s}_{\text{i2v-b}}$), reflecting a balance between temporal coherence and input fidelity. The formulation is given by:

$$S_{\text{ext}} = \frac{\hat{s}_{\text{bg}} + \hat{s}_{\text{smooth}} + \hat{s}_{\text{subj}} + 0.5 \cdot \hat{s}_{\text{i2v-s}} + 0.5 \cdot \hat{s}_{\text{i2v-b}}}{4} \tag{2}$$

# B. Additional Experiments

**Dense Features Outperform Sparse Trajectories.** To ensure a fair comparison within the DiT-based CogVideoX [34] architecture, we align the experimental implementation by adapting the Track4Gen [13] projector to mirror our design: specifically, employing an interpolation layer followed by three MLP layers. Furthermore, we exclude the refiner module to strictly isolate the efficacy of the distillation objective. As shown in Table 3, our method outperforms Track4Gen*. This demonstrates the superiority of dense SAM2 [23] features over sparse point trajectories, as the latter are prone to performance degradation caused by error accumulation in optical flow estimation (e.g., RAFT [29]).

**Explicit Backward Constraints are Redundant.** We investigate the impact of extending the Local Gram Feature (LGF) loss to incorporate a backward temporal constraint (i.e., computing similarity between frame $t$ and $t-1$). As shown in Table 4, this formulation yields inferior results compared to our forward-only design. We attribute this to the inherent bidirectionality

| Configurations | Motion Score↑ | Ext. Motion Score↑ |
|---|---|---|
| Bidirectional LGF | 94.87 | 95.52 |
| **LGF loss (Ours)** | **95.51** | **96.03** |
| 23rd Transformer Block | 94.93 | 95.54 |
| 24th Transformer Block | 94.44 | 95.14 |
| **25th Transformer Block (Ours)** | **95.51** | **96.03** |
| 26th Transformer Block | 94.44 | 95.12 |
| 27th Transformer Block | 94.68 | 95.36 |

Table 4. Ablations on Temporal Directionality and Injection Depth. We analyze two key design choices: (a) the directionality of the Local Gram Feature loss, and (b) the optimal depth for feature injection within the DiT architecture. The highlighted row indicates our default configuration.

of the distilled SAM 2 features; explicitly enforcing backward consistency creates computational redundancy and introduces over-constraints that hamper the generation dynamics. Therefore, we retain the forward-only formulation.

## C. Additional Qualitative Results

We provide additional visual comparisons to further substantiate the effectiveness of `SAM2VideoX`. Specifically, we contrast our method with competing baselines and state-of-the-art models, highlighting superior structural fidelity and temporal coherence across diverse motion scenarios.

## D. Theoretical Analysis of Feature Fusion

**Definition of $\mathrm{S}(\cdot)$.** As referenced in Sec. 4.4, $\mathrm{S}(\cdot)$ denotes a temperature-scaled softmax operation applied to each similarity vector, normalizing the $7 \times 7$ similarity scores into a probability distribution. Given similarities $\{z_i\}_{i=1}^{K}$, we compute

$$p_i = \frac{\exp\left(z_i/T\right)}{\sum_{j=1}^{K} \exp\left(z_j/T\right)},$$

and in all experiments we set the temperature to $T = 0.1$.

**Interference in Feature-Space Fusion.** Let $a = F^{\mathrm{fwd}}_{\mathrm{mem,t}}$, $b = F^{\mathrm{bwd}}_{\mathrm{mem,t}}$, $c = F^{\mathrm{fwd}}_{\mathrm{mem,t+1}}$, and $d = F^{\mathrm{bwd}}_{\mathrm{mem,t+1}}$ represent directional SAM2 memory features across two consecutive frames. The ideal local Gram similarities for the forward and backward teachers are $a \cdot c$ and $b \cdot d$, respectively. If we first fuse forward and backward features in the feature space,

$$f_t = ka + (1-k)b, \quad f_{t+1} = kc + (1-k)d,$$

and then compute the local Gram, we obtain

$$g_{\text{feat}} = f_t \cdot f_{t+1} = k^2(a \cdot c) + (1-k)^2(b \cdot d) + k(1-k)\big(a \cdot d + b \cdot c\big).$$

The final term introduces cross-correlations ($a \cdot d$ and $b \cdot c$) that couple forward features at frame $t$ with backward features at frame $t{+}1$. These cross terms lack a counterpart in the valid teacher similarity matrix (neither purely forward nor backward), thereby mixing incompatible temporal contexts. Consequently, the student model is supervised to learn spurious correlations absent in the teacher's distribution, leading to temporal inconsistencies.

In contrast, our LGF fusion computes local Gram similarities for each direction separately and fuses them only in the LGF space:

$$g_{\text{lgf}} = k(a \cdot c) + (1-k)(b \cdot d).$$

This formulation eliminates cross terms, yielding a convex combination of two consistent teacher signals and circumventing the interference inherent in feature-space fusion.

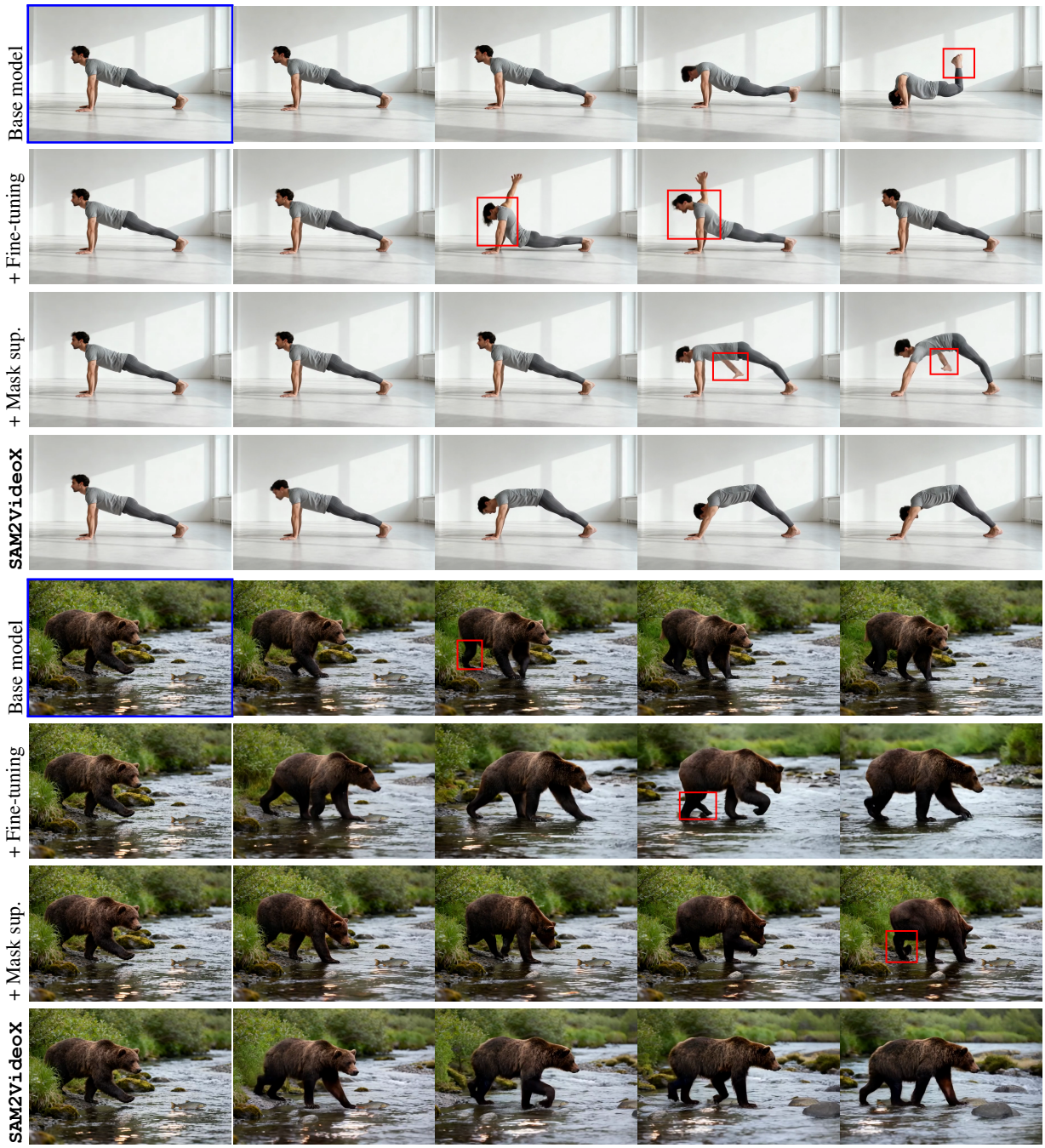


Figure 8. **Extended Visual Comparisons with Baselines.** The blue box indicates the conditioning input image, while red boxes highlight structural failure modes observed in baseline methods.

## E. Limitations and Future Work

**Limitations.** We acknowledge that our model's performance is effectively bounded by the inherent capabilities of the underlying backbone, CogVideoX. specifically, in scenarios involving high-dynamic or complex motion—such as fast-paced dancing or competitive sports—we observe that generation artifacts may remain pronounced. This suggests that while our method significantly improves structural alignment, the ultimate video quality in extreme motion cases relies on the generative capacity of the base model.


Figure 9. **Visual Comparison with State-of-the-Art Open-Source Models.** Notably, despite HunyuanVid [16] possessing 13B parameters, our `SAM2VideoX` (5B) achieves comparable generation quality.

**Future Work.** Our current pipeline relies on SAM 2 for video object segmentation and tracking. While the model demonstrates robust performance for single objects, we observe performance degradation in multi-object scenarios. Specifically, when prompted with visual cues (e.g., bounding boxes or points) for multiple distinct entities, the tracker struggles to maintain consistent identities across temporal sequences. Consequently, exploring effective feature representations for multi-object video generation remains an open and promising direction for future research.

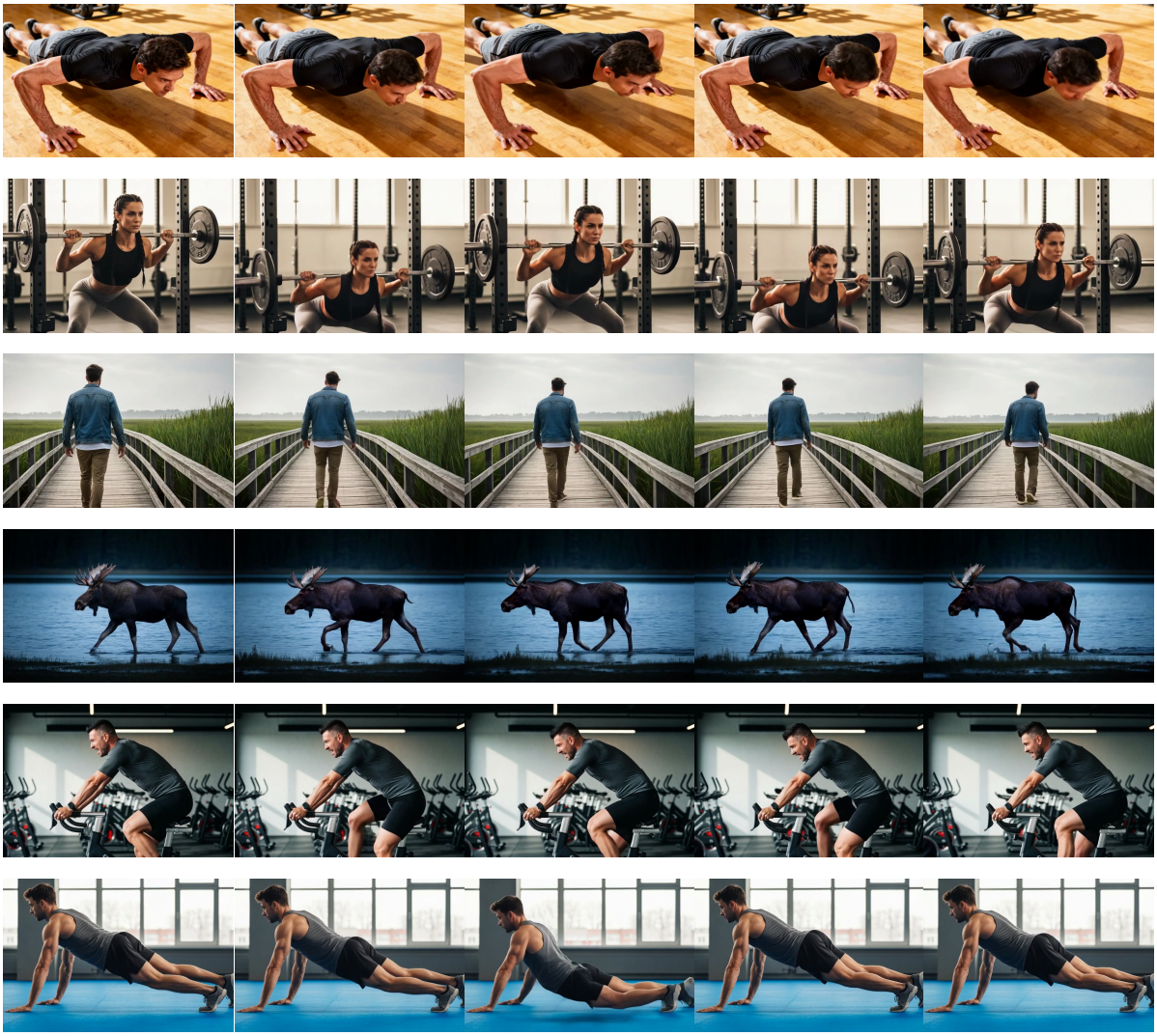

Figure 10. **Additional Qualitative Results.** Our SAM2VideoX maintains robust structural consistency across a diverse range of motion dynamics.