STRUCTURE FROM TRACKING: DISTILLING STRUCTURE-PRESERVING MOTION FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

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 bidirectional video diffusion models. 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.

1 INTRODUCTION

A long philosophical tradition, from Heraclitus to Bergson, casts reality as *becoming* rather than *being* (Heraclitus, 1957; Bergson, 1911). While image generation models have excelled at rendering what is *there* with high fidelity (Rombach et al., 2022; Li et al., 2024), even the most advanced video generation systems (Brooks et al., 2024; Wan et al., 2025; Kong et al., 2024) still struggle to express the true dynamics of *change*.

A persistent 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. Such failures expose a fundamental gap between rendering convincing frames and modeling coherent dynamics. The problem is especially pronounced for articulated and deformable objects such as humans and animals, whose motion requires multiple parts to move coherently while preserving their relative structure.

Existing approaches supply this motion through explicit proxies. At inference time, ControlNet-style conditioning (Zhang et al., 2023) imposes a motion signal but requires the motion to be specified *a priori*, and while rigid motion can often be described with sparse trajectories (Deng et al., 2024), articulated motion is harder to reduce to any single explicit representation: skeletons are category-specific and sparse, while dense displacement fields do not encode part connectivity or identity. At training time, supervision is introduced through optical flow (Chefer et al., 2025), trajectories (Jeong et al., 2024), skeletons, or masks, each of which compresses motion into a predefined abstraction: optical flow emphasizes local displacement, skeletons retain sparse joint geometry, and masks capture object support while discarding most internal part dynamics. When estimated automatically, these proxies inherit errors from off-the-shelf predictors such as RAFT (Teed & Deng, 2020).

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.

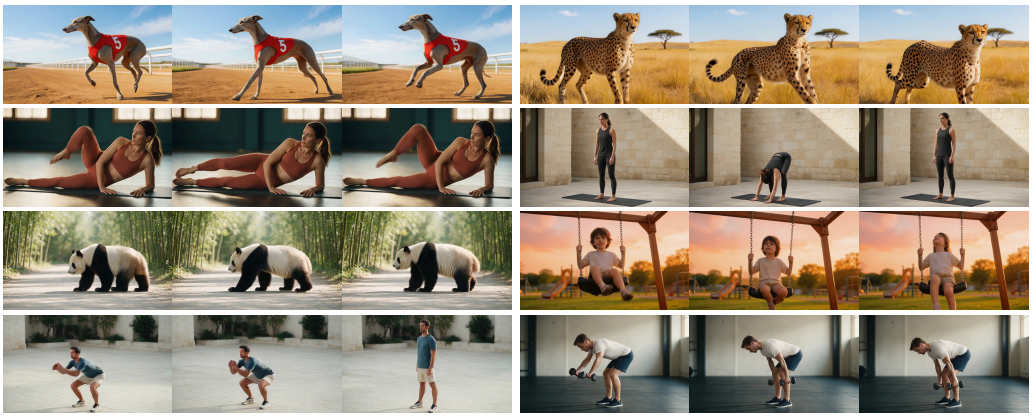

Figure 1: **SAM2VideoX distills structure-preserving motion priors from SAM2 into video diffusion models.** The resulting model generates coherent motion across diverse human and animal subjects while maintaining their structural integrity over time.

An appealing alternative is to distill motion knowledge from a pretrained representation during training, so that the generator acquires a motion prior intrinsically and needs no additional signal at inference. This shifts the difficulty onto the choice of teacher: existing representation-alignment methods (Yu et al., 2024) adopt encoders selected for semantic quality, and whether such representations encode motion at all remains untested. This raises a more basic question:

Rather than hand-designing motion signals,
which pretrained representations best capture structure-preserving dynamics?

We therefore characterize the motion content of pretrained representations learned under objectives that differ in how much temporal structure they require: image self-supervision (DINOv3 (Siméoni et al., 2025)), masked video modeling (VideoMAE V2 (Wang et al., 2023)), predictive video modeling (V-JEPA 2 (Assran et al., 2025)), and video tracking (SAM2 (Ravi et al., 2024)). We quantify this using U-centered motion CKA, which measures whether temporal intervals with similar optical flow induce similar changes in the representation. Across the four encoders, SAM2 exhibits the strongest motion alignment. Crucially, this alignment drops when SAM2 features are paired with motion from a different video or from mismatched temporal locations, indicating that the measured alignment reflects video-specific dynamics rather than generic temporal variation.

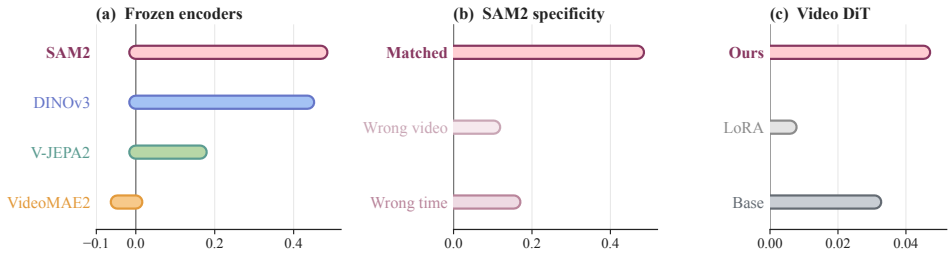

Figure 2: **SAM2 representations are motion-specific and transfer to Video DiT.** (a) Among frozen encoders, SAM2 features show the strongest U-centered motion CKA. (b) Alignment drops under cross-video and temporal mismatch controls, indicating video-specific motion encoding. (c) SAM2-guided distillation increases motion alignment in Video DiT features beyond the base model and matched LoRA. Full details are provided in Appendix A.

These results identify tracking representations, and SAM2 in particular, as a promising source of motion supervision, matching the intuition that successful tracking requires learning how parts move together, how limbs stay connected, and how occlusions resolve over time. We propose `SAM2VideoX`, which distills SAM2's internal memory features into a video diffusion model. Distilling the internal representation rather than the output masks preserves part correspondence and temporal structure that a binary mask discards.

Distilling this knowledge into a generator, however, presents two technical challenges. First, this distillation is hindered by architectural asymmetry: state-of-the-art video generation models use DiT (Peebles & Xie, 2023) architectures with bidirectional attention for global context, while SAM2 is inherently causal and recurrent. To bridge this gap, `SAM2VideoX` extracts SAM2 features in both forward and reversed temporal orders and fuses them into a bidirectional teacher representation with access to both past and future context. Second, direct feature regression assumes a meaningful coordinate-wise correspondence between heterogeneous teacher and student feature spaces, an unnecessarily restrictive requirement for transferring motion structure. We therefore introduce a *Local Gram Flow* (LGF) loss that aligns local cross-frame feature relations rather than absolute feature coordinates. This supervision produces a measurable change in the generator itself: after distillation, intermediate Video DiT features exhibit substantially stronger motion alignment than the pretrained model and a step-matched LoRA control (Fig. 2c).

Across qualitative and quantitative evaluations, `SAM2VideoX` yields more realistic, structurally coherent motion. On our VBench-I2V evaluation (Huang et al., 2024; 2025), `SAM2VideoX` achieves a Motion Score of 95.51, surpassing REPA (Yu et al., 2024) by 2.60 points. Our FVD is 360.57, a reduction of 21.20% and 22.46% versus REPA- and LoRA-fine-tuning (Hu et al., 2022). Because existing benchmarks only partially capture structural failures in articulated motion, we additionally conduct a human evaluation, where `SAM2VideoX` is preferred in 71.4% of pairwise judgments across 48 participants under a double-blind forced-choice protocol.

2 Related Work

Video diffusion models. Video diffusion models have progressed rapidly, from UNet-based models such as Stable Video Diffusion (Blattmann et al., 2023) to DiT-based models including CogVideoX (Yang et al., 2024), HunyuanVid (Kong et al., 2024), Wan-Video (Wan et al., 2025), and Cosmos (NVIDIA et al., 2025). Despite strong visual fidelity, generating coherent structure-preserving motion remains challenging, especially for articulated and deformable objects. Our goal is to improve this intrinsic capability without requiring auxiliary motion controls at inference.

Motion understanding. Understanding motion has long been central to video analysis (Horn & Schunck, 1981; Perazzi et al., 2016). Optical flow and point trajectories (Teed & Deng, 2020; Karaev et al., 2024) provide explicit motion correspondences, but carry limited object-level structural semantics and can become unreliable under challenging motion and occlusions. SAM2 (Ravi et al., 2024) tracks objects across long sequences while maintaining object identity through occlusions. Rather than using its discrete, boundary-focused output masks, we leverage SAM2's dense and temporally consistent internal features as motion priors for structure preservation.

Representation alignment. Representation alignment was introduced for image generation by REPA (Yu et al., 2024) and has recently been extended to video generation. Existing approaches align diffusion features with explicit motion signals such as trajectories or flow (Jeong et al., 2024; Chefer et al., 2025), or with pretrained visual representations, as in VideoREPA (Zhang et al., 2025), CREPA (Hwang et al., 2025), and MoAlign (Bhowmik et al., 2026). In contrast, we align diffusion features with SAM2's tracking representations and introduce bidirectional fusion and Local Gram Flow to transfer their structure-preserving motion priors.

3 Preliminaries

We briefly review latent video diffusion and SAM2, which serve as the student generator and motion teacher in our framework, respectively.

Latent Video Diffusion Model. We adopt a latent video diffusion model with a DiT denoiser (Peebles & Xie, 2023). Given a video $\mathbf{x}$, a pretrained VAE encoder $\mathcal{E}$ maps it to a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$. At diffusion timestep t, the latent is perturbed as $\mathbf{z}_t = \alpha_t \mathbf{z} + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The denoiser f_θ is trained with the velocity-prediction objective $\mathcal{L}_{\text{diff}}$, with target $\mathbf{v} = \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{z}$. We use its intermediate activations as the student representation for motion distillation.

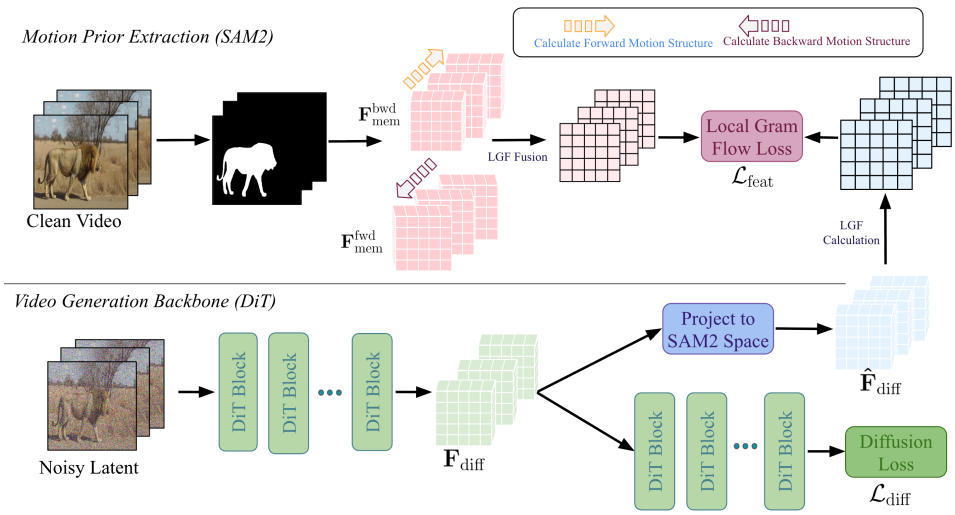

Figure 3: **Method overview.** The framework consists of two parallel branches: **(Top)** The *Motion Prior Extraction* branch extracts forward and backward memory features from SAM2 given a clean video, and fuses them into a bidirectional teacher representation. **(Bottom)** The *Video Generation Backbone* 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** is used to align the spatio-temporal structure of the projected student features with the teacher priors.

Segment Anything Model 2 (SAM2). The Segment-Anything Model 2 (SAM2) extends the image segmentation capability of SAM 1 (Kirillov et al., 2023) 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 a 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_{mem}. A mask decoder $\mathcal{D}_{\text{mask}}$ subsequently takes F_{mem} and user prompts to generate the segmentation mask. By applying this recurrent procedure across all frames, SAM2 produces both a sequence of masks and a sequence of memory features:

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

4 Method

We distill SAM2's structure-preserving motion priors into a video DiT through internal feature alignment. We project intermediate DiT features for relational matching (Sec. 4.1), use Local Gram Flow (LGF) to align local cross-frame relations (Sec. 4.2), and fuse forward and backward SAM2 relations in LGF space to construct a bidirectional teacher (Sec. 4.3). The resulting LGF distillation loss is optimized jointly with the diffusion objective (Sec. 4.4). Fig. 3 summarizes the framework.

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_θ, 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. As analyzed in Fig. 2 and Appendix A, SAM2 memory representations exhibit substantially stronger alignment with video-specific motion than generic visual representations. We therefore use them as the distillation target.

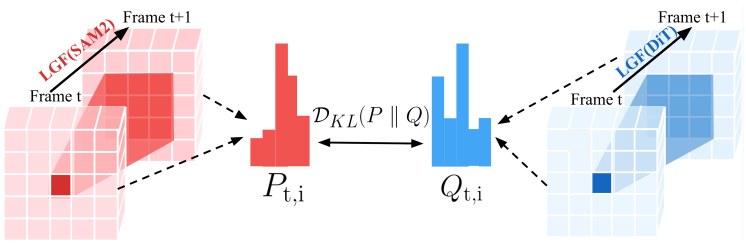

Figure 4: **Illustration of the Local Gram Flow (LGF) Loss.** The LGF operator captures motion structure by computing similarities between a token at frame t and its spatial neighborhood in the subsequent frame $t+1$. Instead of matching absolute values (e.g., ℓ_2), we convert the resulting similarity vectors $(P_{\text{t,i}}, Q_{\text{t,i}})$ into probability distributions and align them using KL divergence. This forces the model to learn relative motion patterns, not just feature values.

To align the DiT representation $\mathbf{F}_{\text{diff}}$ with the SAM2 teacher, we introduce a lightweight projector $\mathcal{P}$, consisting of temporal resampling followed by an MLP, which maps the DiT activations to the spatio-temporal resolution and channel dimension required for relational matching:

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

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 match their *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 spatial neighborhood in the subsequent frame (see Fig. 4). This produces, at each spatial position, a local cross-frame similarity vector that encodes where its feature is most likely to correspond in the next frame. We denote this operator as $\text{LGF}(\cdot)$.

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 forward and backward passes 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:

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

Empirically, using separate projectors to align separately to $\mathbf{F}^{\text{fwd}}_{\text{mem}}$ and $\mathbf{F}^{\text{bwd}}_{\text{mem}}$ 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 demonstrates (Fig. 8), naively adding them $(k\mathbf{F}^{\text{fwd}}_{\text{mem}} + (1-k)\mathbf{F}^{\text{bwd}}_{\text{mem}})$ leads to severe performance degradation. Because the

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. Higher Motion/Extended Motion Scores indicate better structure preservation; lower FVD indicates superior perceptual quality.

Method	BC↑	SC↑	MS↑	Motion Score↑	Ext. Motion Score↑	FVD↓
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
+ Point Traj. Supervision	97.32	94.67	98.35	95.11	<u>95.69</u>	369.33
+ Mask Supervision	<u>97.86</u>	94.43	<u>98.42</u>	<u>95.34</u>	95.54	397.73
+ REPA	97.41	91.99	97.31	92.91	93.77	457.59
+ VideoREPA	97.71	94.24	98.23	94.98	95.52	<u>364.82</u>
+ V-JEPA 2 Supervision	97.23	**94.88**	98.34	95.15	95.44	369.09
+ **SAM2VideoX (Ours)**	**97.88**	<u>94.76</u>	**98.45**	**95.51**	**96.03**	**360.57**
HunyuanVid	96.85	95.32	98.76	95.62	96.24	583.99

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:

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

4.4 Motion Distillation Loss

Finally, the motion distillation loss, $\mathcal{L}_{\mathrm{feat}}$, is designed to match the LGF distributions of the student $\hat{\mathbf{F}}_{\mathrm{diff}}$ and the fused teacher $\mathbf{F}_{\mathrm{SAM2}}$, 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:

$$\mathbf{P} = \mathrm{Softmax}\Big(\ \mathrm{LGF}(\mathbf{F}_{\mathrm{SAM2}})\Big),$$
$$\mathbf{Q} = \mathrm{Softmax}\Big(\ \mathrm{LGF}(\hat{\mathbf{F}}_{\mathrm{diff}})\Big).$$

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

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

Our ablation studies (Fig. 8) 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 9,837 single-subject motion clips from Panda70M (Chen et al., 2024), MMTrail (Chi et al., 2024), and MotionVid (Wang et al., 2025), covering diverse human and animal motions. We instantiate `SAM2VideoX` on two video diffusion backbones, CogVideoX-5B-I2V (Yang et al., 2024) and Wan2.2-5B (Wan et al., 2025). Unless otherwise specified, quantitative comparisons, teacher studies, ablations, and the human evaluation use the CogVideoX-based instantiation; our main qualitative comparison in Fig. 5 uses Wan2.2-5B. Backbone-specific training and implementation details are provided in Appendix A.

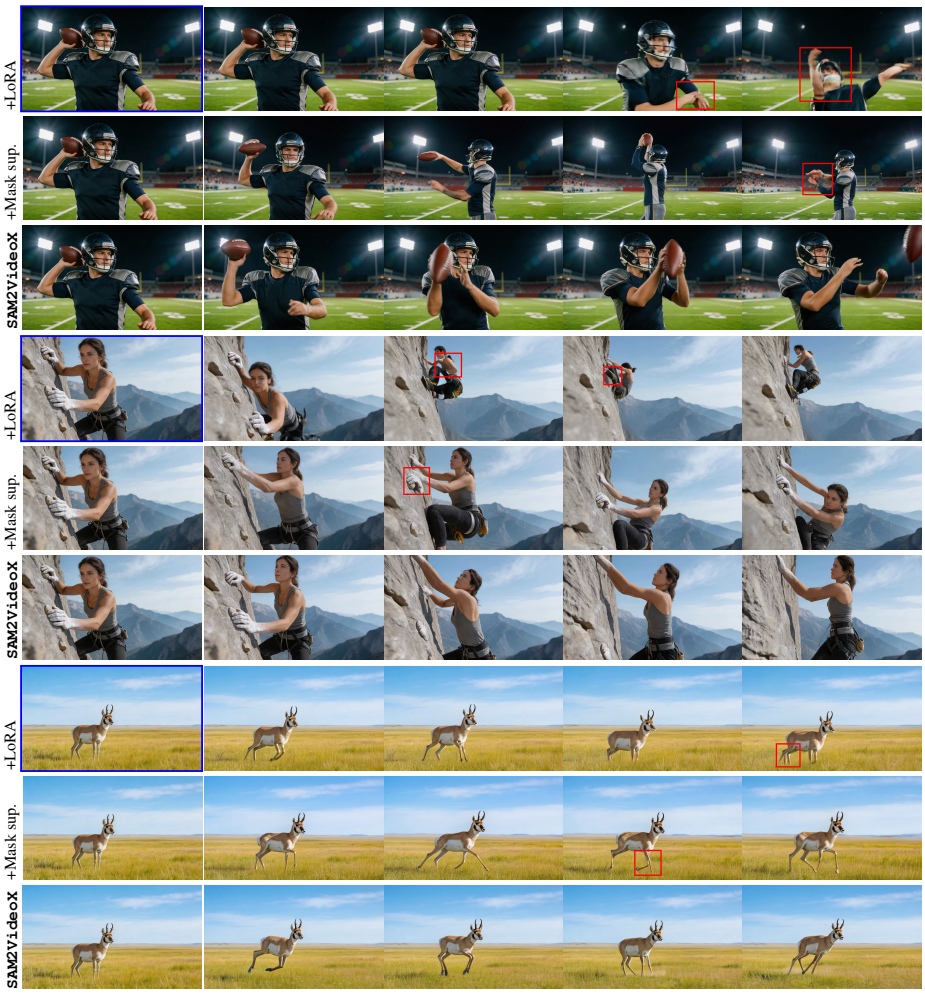

Figure 5: **Qualitative comparison on articulated motion using Wan2.2-5B.** The blue box indicates the input image. Red boxes highlight structural failures in the baselines, including distorted limbs and inconsistent articulated motion. In contrast, `SAM2VideoX` better preserves structural integrity throughout the generated sequences. "+LoRA" denotes LoRA fine-tuning and "+Mask sup." denotes mask supervision. We recommend viewing the supplementary videos.

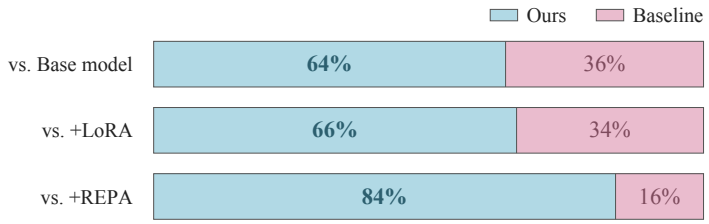

Figure 6: **Human preference win rates.** In pairwise comparisons of motion realism and physical plausibility, our method was preferred ($\geq 64\%$) over all baselines.

Baselines. Unless otherwise specified, all fine-tuned variants share the same base model and are trained on our curated dataset under matched LoRA and optimization settings. We include the unmodified **base model** and **LoRA fine-tuning** with diffusion loss only, and organize the remaining baselines according to the design choice they isolate.

Supervision granularity. **+ Mask Supervision** attaches a linear head to our projection head $\mathcal{P}$ to predict the subject mask, replacing feature alignment with mask supervision. **+ Point Traj. Supervision** uses sparse point-tracking correspondences following Track4Gen (Jeong et al., 2024).

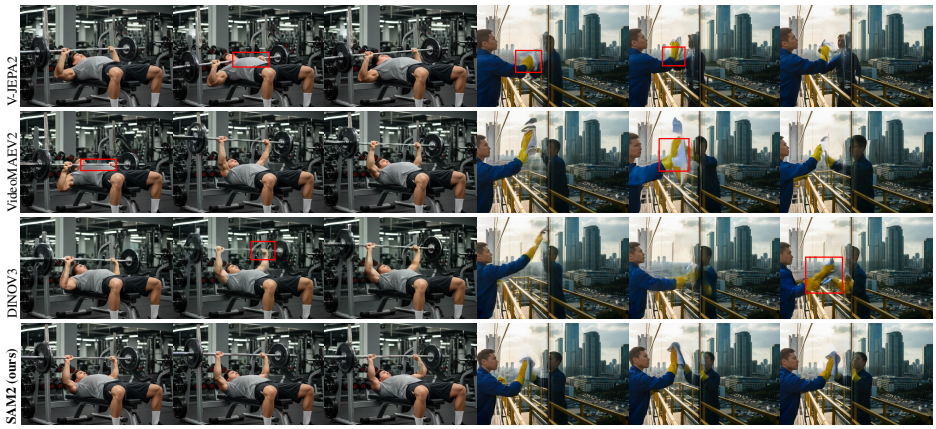

Figure 7: **Qualitative comparison on different teacher representations.** Red boxes highlight structural artifacts produced using V-JEPA 2, VideoMAE V2, and DINOv3 as distillation teachers. In contrast, SAM2 supervision better preserves articulated structure and produces more coherent motion, supporting the benefit of tracking-oriented representations for motion distillation.

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+Backward LGF	94.87	95.52
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**

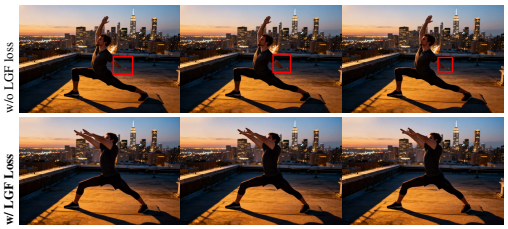

Figure 8: **Ablation of core components. Left:** Quantitative ablations show that LGF-KL and bidirectional LGF fusion consistently outperform their alternatives. **Right:** Qualitatively, raw-feature ℓ_2 alignment produces temporal artifacts, whereas LGF-KL yields smoother and more stable motion.

Together, these baselines test whether explicit region- or point-level supervision can replace dense feature distillation.

Teacher representation. **+ REPA** (Yu et al., 2024) aligns DiT features with the image-based DINOv3 encoder (Siméoni et al., 2025). **+ VideoREPA** (Zhang et al., 2025) instead uses VideoMAE V2 features learned through masked video modeling, while **+ V-JEPA 2 Supervision** (Assran et al., 2025) uses representations learned through predictive video modeling. These comparisons distinguish the benefit of generic image and video representations from the tracking-oriented memory representations of SAM2.

Evaluation protocol. We evaluate along three axes. *(i) Motion.* On 85 prompts from VBench-I2V (Huang et al., 2024; 2025), we report Motion Smoothness, Subject Consistency, and Background Consistency, and summarize them as a Motion Score (their mean after min-max normalization using the official VBench-I2V protocol); the Extended Motion Score additionally includes I2V-Subject and I2V-Background Consistency with weight 0.5. *(ii) Perceptual quality.* FVD (Unterthiner et al., 2018) on 200 held-out 49-frame clips from the training dataset, where the evaluated clips are excluded from training although other temporal crops from the same source videos may be used. *(iii) Human preference.* A double-blind study on 40 prompts in which 48 participants choose between two side-by-side videos (ours vs. baseline, order randomized). For the default quantitative and human evaluations, all methods generate 49-frame 720×480 videos at 8 fps, with 50 denoising steps and guidance scale 6.0. Wan2.2-5B evaluation settings are provided in Appendix A.

5.1 Results

SAM2VideoX improves both structure preservation and perceptual quality. Table 1 shows that SAM2VideoX consistently improves over the base model and fine-tuning baselines, achieving the best results on five of six metrics among CogVideoX-based methods. It attains the highest Motion Score (95.51) and Extended Motion Score (96.03), while reducing FVD to 360.57. These results show that SAM2 distillation improves both temporal coherence and perceptual quality.

Feature distillation outperforms explicit structural supervision. We compare our feature-level objective with explicit mask and point-trajectory supervision under the same training setup. While both explicit targets improve over LoRA fine-tuning, SAM2 feature distillation performs best overall, suggesting that its dense intermediate representations retain motion structure not fully captured by region-level masks or sparse trajectories. Additional qualitative comparisons on the same base model in Appendix Fig. 9 show the same trend. This advantage also transfers to the stronger Wan2.2-5B backbone, where Fig. 5 shows better preservation of articulated structure than LoRA fine-tuning and mask supervision.

Tracking-oriented representations provide stronger motion supervision. To examine how the choice of teacher representation affects video generation, we compare against representation-alignment baselines using both image and video encoders. REPA (Yu et al., 2024), which aligns DiT features with image-based DINOv3 representations, achieves a Motion Score of 92.91. Using video representations improves motion consistency: VideoREPA and V-JEPA 2 supervision reach Motion Scores of 94.98 and 95.15, respectively. Nevertheless, `SAM2VideoX` achieves the strongest Motion Score (95.51), Extended Motion Score (96.03), and FVD (360.57) among these representation-alignment baselines. These results suggest that SAM2's tracking-oriented features provide a particularly effective supervisory signal for structure-preserving motion.

Human evaluators consistently prefer our method. As shown in Fig. 6, `SAM2VideoX` is preferred over the base model and all evaluated fine-tuning baselines in our double-blind study. This indicates that the improvements in automatic metrics correspond to perceptually meaningful gains in motion realism and structural coherence.

5.2 Ablations

We conduct systematic ablations on the 85-prompt VBench-I2V subset to isolate the contributions of our two core designs: bidirectional LGF fusion and the LGF-KL distillation objective.

LGF-space fusion is critical for bidirectional supervision. Using only the forward SAM2 stream ($\mathbf{F}^{\text{fwd}}_{\text{mem}}$) already provides a strong teacher, achieving a Motion Score of 95.07. However, naively incorporating the backward stream does not yield further gains: using separate projectors reduces the score to 94.68, while directly averaging forward and backward features performs only marginally above LoRA fine-tuning (94.16 vs. 94.02). These results show that simply combining opposite temporal directions at the raw-feature level is ineffective. In contrast, fusing the two directional teachers in LGF space achieves the best Motion Score of 95.51. We further test whether LGF itself should be temporally symmetric by adding an explicit $t \to t-1$ constraint to the default $t \to t+1$ relation. This forward+backward variant reduces the Motion Score to 94.87, suggesting that the bidirectional SAM2 teacher already provides sufficient two-sided temporal context and that an additional backward LGF constraint is redundant. Qualitative results in Fig. 12 further show that LGF-space fusion better preserves articulated structure.

KL divergence outperforms ℓ_2 for structural alignment. We next ablate the distillation objective while keeping the teacher fixed. Direct ℓ_2 matching of raw features achieves a Motion Score of 94.58, while replacing KL with ℓ_2 in the LGF space performs similarly at 94.51. Both are substantially below the full LGF-KL objective at 95.51. These results show that neither direct feature regression nor relational matching alone is sufficient: LGF captures local cross-frame structure, while KL aligns the resulting similarity distributions without enforcing exact value correspondence. Their combination provides the strongest supervision for structure-preserving motion. Fig. 8 further shows that LGF-KL reduces temporal artifacts and yields smoother motion.

6 Conclusion

We present SAM2VideoX, which distills structure-preserving motion priors from SAM2 into video diffusion models. Guided by our finding that SAM2 representations exhibit strong video-specific motion alignment, we introduce bidirectional teacher fusion and Local Gram Flow alignment to transfer these priors into the generator. Experiments on CogVideoX and Wan2.2 show consistent gains in motion coherence and structural preservation without additional inference-time controls. More broadly, our findings highlight tracking representations as a valuable source of transferable motion priors for improving the intrinsic dynamics of video generation models.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing and organization. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

Our evaluation includes a human preference study in which participants provided forced-choice judgments between generated videos. The study did not require participants to provide sensitive personal information. As with video generation systems more broadly, improvements in motion realism may increase the potential for misuse of synthetic media.

REPRODUCIBILITY STATEMENT

We provide the model architecture, feature extraction procedure, optimization settings, and implementation details in Sec. 4–5 and Appendix A. Appendix B describes the evaluation protocol and metric definitions, Appendix C and D provide additional ablations and analysis details, and Appendix G lists all VBench-I2V prompts used in evaluation. Anonymized reference code for the core components on Wan2.2-5B is included in the supplementary material.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Henri Bergson. *Creative Evolution*. Henry Holt and Company, New York, 1911. English translation of *L'Evolution créatrice* (1907).

Aritra Bhowmik, Denis Korzhenkov, Cees G. M. Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. In *International Conference on Learning Representations*, 2026.

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.

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. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

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.

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.

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. URL https://arxiv.org/abs/2407.20962.

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*, pp. 183–199. Springer, 2024.

Heraclitus. Fragments. In G. S. Kirk and J. E. Raven (eds.), *The Presocratic Philosophers*. Cambridge University Press, Cambridge, 1957. Ancient fragments conventionally dated to ca. 500 BCE.

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

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

Sungwon Hwang, Hyojin Jang, Kinam Kim, Minho Park, and Jaegul Choo. Cross-frame representation alignment for fine-tuning video diffusion models, 2025. URL https://arxiv.org/abs/2506.09229.

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.

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.

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*, pp. 4015–4026, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

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.

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*, pp. 38–55. Springer, 2024.

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. *arXiv preprint arXiv:2511.00062*, 2025.

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.

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*, pp. 724–732, 2016.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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.

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. *arXiv preprint arXiv:2508.10104*, 2025.

Gábor J. Székely and Maria L. Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014. doi: 10.1214/14-AOS1255.

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*, pp. 402–419. Springer, 2020.

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.

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.

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*, pp. 12391–12401, 2025.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, June 2023.

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.

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.

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.

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.

A Implementation Details

Feature Analysis Details. To analyze how strongly different pretrained representations encode video motion, we compare frozen features from DINOv3, VideoMAE V2, V-JEPA 2, and SAM2 on a fixed set of 200 human and animal videos. Each video contains 121 frames at a resolution of 832×480. We extract spatial feature maps from each frozen encoder before applying any learned analysis head. Since the encoders operate at different spatial and temporal resolutions, we temporally resample their features to a common sequence and spatially pool them onto the same grid over the shared visible region.

To make representations comparable across encoders with different feature spaces, we characterize each frame by the pairwise similarities among its spatial features. Let $\mathbf{F}_t$ denote the spatial feature map at time t. After normalizing the feature vectors, we compute

$$\mathbf{G}_t = \hat{\mathbf{F}}_t \hat{\mathbf{F}}_t^\top,$$

and represent temporal feature change by

$$\Delta\mathbf{G}_t = \mathbf{G}_{t+1} - \mathbf{G}_t.$$

As a common reference for video motion, we compute optical flow over the corresponding temporal intervals and spatially pool it onto the same grid.

We measure the alignment between the sequence of representation changes and the corresponding motion sequence using U-centered CKA (Kornblith et al., 2019; Székely & Rizzo, 2014). Specifically, we construct temporal Gram matrices $\mathbf{K}$ and $\mathbf{L}$ from the representation changes and motion features, respectively, apply standard U-centering $\mathcal{U}(\cdot)$, and compute

$$\mathrm{CKA}_{\mathrm{motion}} = \frac{\langle \mathcal{U}(\mathbf{K}), \mathcal{U}(\mathbf{L}) \rangle_F}{\|\mathcal{U}(\mathbf{K})\|_F \|\mathcal{U}(\mathbf{L})\|_F}.$$

A higher score indicates that temporal changes in the representation more closely follow the motion dynamics of the underlying video.

To verify that this alignment reflects video-specific motion rather than generic temporal variation, we additionally evaluate two controls for SAM2. The cross-video control pairs representation changes with motion from a different video, while the temporal-mismatch control pairs them with motion from mismatched temporal locations within the same video.

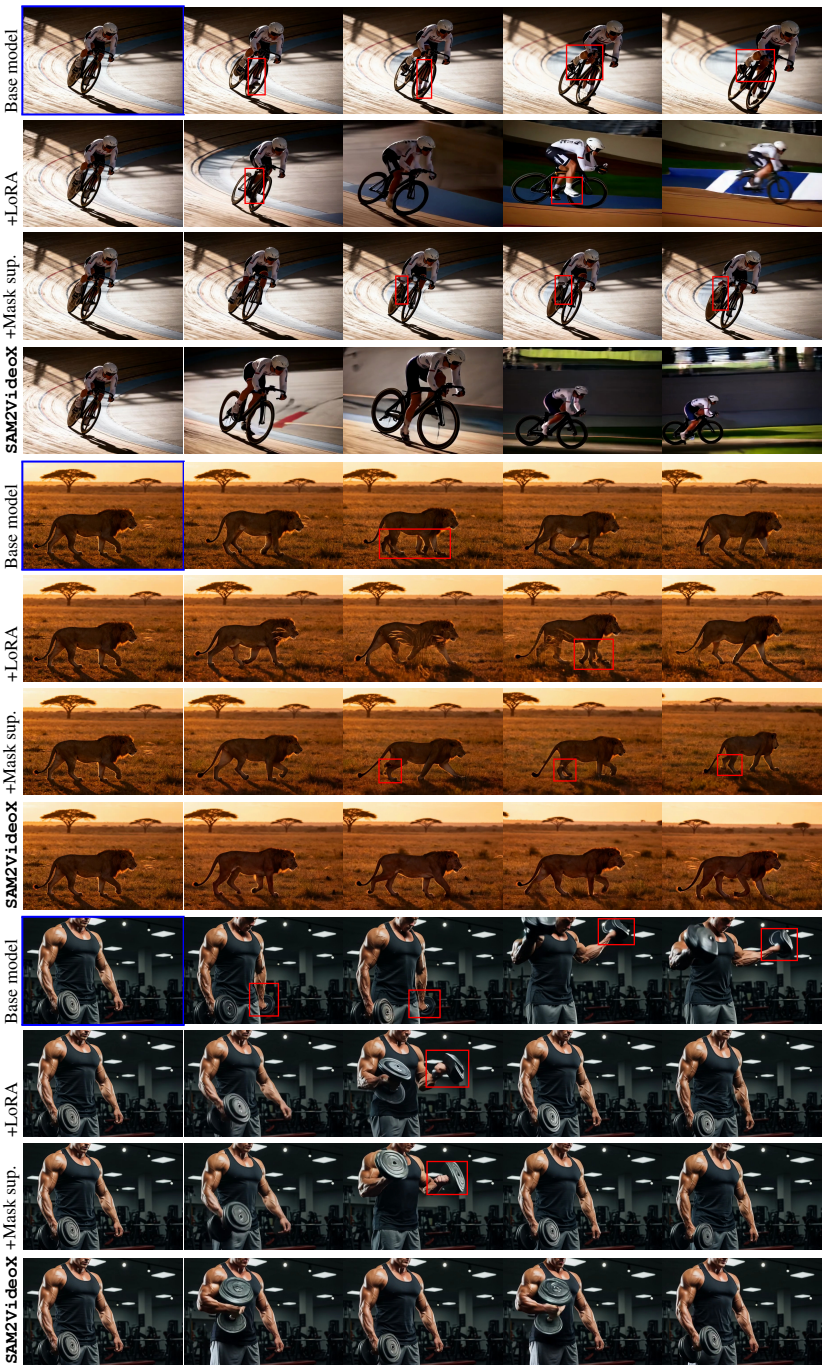

Figure 9: **Qualitative comparison on articulated motion.** The blue box indicates the input image. Red boxes highlight common failure modes in baselines, including structural distortion (cyclist's legs) and physical implausibility (three-legged lion, inconsistent leg motion). In contrast, our method (`SAM2VideoX`) consistently maintains structural integrity. "+LoRA" is the LoRA fine-tuning baseline; "+ Mask sup." is the mask supervision baseline.

For the Video DiT analysis in Fig. 2(c), we use Wan2.2-TI2V-5B and extract the raw intermediate features from its 18th DiT block at a fixed diffusion timestep of $t = 0.75$. We use the same video inputs, noise realizations, timestep, and feature extraction layer for the base model, the matched LoRA baseline, and our model.

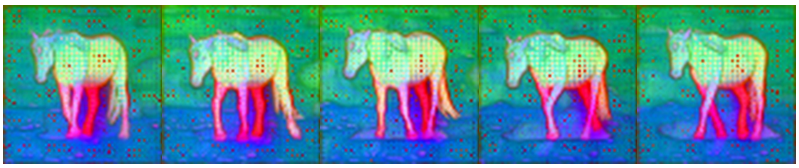

Figure 10: **PCA Visualization of SAM2 Tracking Priors.** PCA components reveal consistent part-level correspondences across frames, indicating that SAM2 memory features capture fine-grained structural dynamics.

We also visualize the SAM2 memory feature $\mathbf{F}_{\text{SAM2}}$ via PCA (Fig. 10), which reveals consistent part-level correspondences across frames.

Feature Projector 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 use a temporal interpolation factor of 4, accompanied by a skip connection with a kernel size of $(3, 1, 1)$ and 768 channels. The subsequent MLP layer widths follow the sequence $768 \to 512 \to 256 \to 256$. For the LGF, we set the spatial local neighborhood size to 7×7.

Table 2: **Quantitative comparison on Wan2.2-5B.** Our method consistently improves motion quality over the fine-tuned baseline and mask supervision, demonstrating its generalizability across video generation architectures.

Method	BC↑	SC↑	MS↑	Motion Score↑	Ext. Motion Score↑
Fine-tuned Wan2.2-5B	<u>97.28</u>	95.13	<u>99.09</u>	96.11	96.35
+ Mask Supervision	97.22	**95.46**	99.05	<u>96.18</u>	<u>96.42</u>
+ **SAM2VideoX (Ours)**	**97.43**	<u>95.28</u>	**99.18**	**96.35**	**96.57**

CogVideoX Training Details. Source videos are sampled at 8 fps and capped at 100 frames, from which we sample 49-frame clips for training. We use GroundingDINO (Liu et al., 2024) to obtain subject bounding boxes for SAM2 prompting. The first-frame mask initializes forward tracking to extract $\mathbf{F}_{\text{mem}}^{\text{fwd}}$, while the last-frame mask initializes backward tracking to extract $\mathbf{F}_{\text{mem}}^{\text{bwd}}$. To avoid SAM2 inference overhead during training, we precompute these features for clips with starting frames spaced at 20-frame intervals. For bidirectional LGF fusion, we set the fusion coefficient to $k = 0.7$.

We apply LoRA (Hu et al., 2022) exclusively to the attention modules with rank 256 and scaling factor $\alpha = 128$, keeping all other backbone parameters frozen. We train for 3,000 steps on $8\times$ H200 GPUs with a global batch size of 32, using gradient accumulation over four steps per GPU. LoRA parameters are optimized with AdamW using $(\beta_1, \beta_2) = (0.9, 0.95)$ and a 200-step linear warmup to a peak learning rate of 10^{-4}. For the projector, we use a cosine decay scheduler with a 150-step warmup; the learning rate is initialized at 5×10^{-4} and decays to 1×10^{-5}. We set the gradient clipping threshold to 1.0 and the feature alignment weight to $\lambda = 0.5$.

Extending to Wan2.2-5B. For Wan2.2-5B, we first employ a $1 \times 1 \times 1$ convolution to project the 3072-channel DiT features to 1024, followed by Group Normalization and SiLU activation. The temporal interpolation factor is 4, accompanied by a residual depthwise-separable temporal convolution with a kernel size of $(3, 1, 1)$. Spatial dimensions are upsampled by a factor of $2\times$ via bilinear interpolation. The subsequent MLP layer widths follow $1024 \to 512 \to 256 \to 256$. For LGF, we use a 3×3 spatial neighborhood.

We apply LoRA (Hu et al., 2022) only to the self-attention modules, with rank 64. We use a 100-step linear warmup, increasing the learning rate from 10^{-8} to a peak of 10^{-4}. For the projector, we use a cosine decay scheduler with a 50-step warmup; the learning rate is initialized at 10^{-3} and decays to 10^{-4}. We set the gradient clipping threshold to 1.0. The model is trained for 600 steps with an effective batch size of 32. At inference, Wan2.2-5B generates 121 frames at 832×480 resolution

Table 3: **Additional VBench metrics and source-video-disjoint FVD.** `SAM2VideoX` substantially improves perceptual quality on 100 held-out source videos while maintaining comparable aesthetic, imaging, and image-conditioning fidelity to the base model. AQ and IQ denote Aesthetic Quality and Imaging Quality.

Method	AQ↑	IQ↑	I2V-SC↑	I2V-BC↑	FVD↓
CogVideoX-I2V	**58.40**	<u>69.72</u>	**96.92**	<u>99.11</u>	648.81
+ LoRA	57.29	68.58	95.83	99.03	535.13
+ REPA	54.54	66.60	94.89	99.04	517.32
+ **Ours**	57.49	**69.83**	<u>96.81</u>	**99.24**	**492.70**

Configurations	Motion Score↑	Ext. Motion Score↑
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 injection depth.** We evaluate the injection depth within the DiT architecture. The highlighted row indicates our default configuration.

and 24 fps. We use the same feature alignment weight $\lambda = 0.5$ and bidirectional fusion coefficient $k = 0.7$ as for CogVideoX.

Quantitative results for baselines implemented on Wan2.2-5B are shown in Table 2, where our method outperforms the alternatives across nearly all metrics.

B 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 select a subset of 85 prompts from the VBench-I2V (Huang et al., 2024) benchmark. The selected prompts predominantly feature human or animal subjects performing articulated motion. We exclude Aesthetic Quality and Imaging Quality from the primary motion composites and report them separately in Table 3. The complete list of selected prompts is provided in Section G.

Metric Selection. Since our primary objective is structure-preserving video generation rather than artistic stylization, we exclude general visual quality metrics such as *Aesthetic Quality* and *Imaging Quality*, which are less indicative of structural fidelity. To provide a unified quantitative evaluation, we define 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, mapping them to a common 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 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 additionally account for fidelity to the conditioning image, the Extended Motion Score incorporates two I2V consistency metrics. Each I2V metric—I2V Subject Consistency ($\hat{s}_{\text{i2v-s}}$) and I2V Background Consistency ($\hat{s}_{\text{i2v-b}}$)—is weighted by 0.5, reflecting a balance between temporal coherence and input fidelity.

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

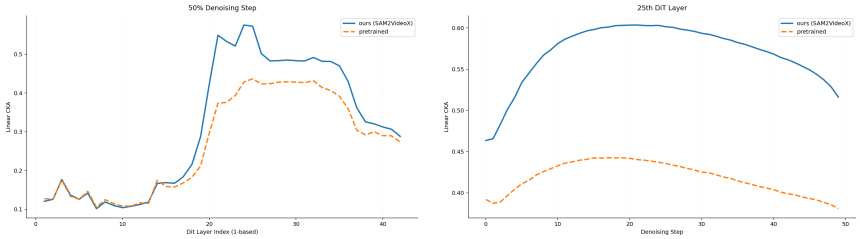

Figure 11: **Verification of Alignment.** Linear CKA similarity between CogVideoX DiT features and SAM2 features, measured before and after distillation with our LGF loss. **Left:** CKA across transformer blocks at a fixed timestep. **Right:** CKA across diffusion timesteps at the 25th block. Higher scores indicate stronger structural alignment.

C Additional Quantitative Results

Additional VBench Metrics and Source-Disjoint FVD. We additionally report Aesthetic Quality (AQ), Imaging Quality (IQ), I2V Subject Consistency (I2V-SC), and I2V Background Consistency (I2V-BC) following the official VBench-I2V protocol. AQ and IQ measure visual quality and I2V-SC/I2V-BC evaluate fidelity to the conditioning image. As an additional robustness check for perceptual quality, we evaluate FVD on 100 source-video-disjoint videos; neither these videos nor any temporal crops from them are used during training. Results are shown in Table 3.

DiT Feature Selection. To determine the optimal injection layer, we compute the Linear Centered Kernel Alignment (CKA) between CogVideoX DiT features and SAM2 features across all transformer blocks. As shown in Fig. 11, the intrinsic alignment peaks in the mid-to-late layers. This identifies the 25th block as the optimal injection point, which balances high-frequency noise in earlier layers against excessive semantic abstraction in later layers. This selection is corroborated by the layer ablation in Table 4. To quantify the distillation effect, we compare CKA scores before and after training. As shown in Fig. 11, training with our LGF loss leads to a substantial increase in CKA scores across all layers. This demonstrates that the LGF loss effectively aligns the DiT's internal representations with the teacher's structure-aware features.

D Analysis of LGF-Space Fusion

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

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

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

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 $d = F^{\mathrm{bwd}}_{\mathrm{mem,t+1}}$ denote the directional SAM2 memory features at two consecutive frames. The ideal local Gram similarities for the forward and backward teachers are $a \cdot c$ and $b \cdot d$, respectively. Consider a naive baseline that first fuses forward and backward features in the feature space:

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

Computing the local Gram similarity of these fused features gives:

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

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 have no counterpart in either teacher's similarity matrix, thereby mixing incompatible temporal contexts. As a result, the student is supervised toward spurious correlations absent from the teacher's distribution, degrading temporal consistency.



Figure 12: **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.

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 clean convex combination of the two teacher signals without the interference introduced by feature-space fusion.

E Additional Qualitative Results

We provide additional qualitative results across diverse motion scenarios. Figure. 12 examines our bidirectional LGF fusion, Figure. 13 compares against diferent baselines. Additional representative generations from `SAM2VideoX` are shown in Figure. 14.

F Limitations and Future Work

Limitations. Our current framework is primarily designed for single-subject motion. Its teacher signal relies on SAM2 maintaining a consistent tracked identity, which can become ambiguous in videos with multiple interacting objects. In addition, LGF models cross-frame correspondences within a fixed local neighborhood, so very large or abrupt displacements may exceed its effective matching range.

Future Work. An important direction is to extend the framework to multi-object and interaction-heavy videos using identity-aware teacher representations. It would also be interesting to develop multi-scale or adaptive relational matching to capture larger displacements while retaining the local structural constraints of LGF. More broadly, our formulation could be extended to distill motion priors from stronger tracking or correspondence models as they become available.

G Selected VBench-I2V Prompts

1. *"A person is whisking eggs, and the egg whites and yolks are gently streaming out"*
2. *"A teddy bear is climbing over a wooden fence"*
3. *"A young boy is lifting a bundle of dry grass wrapped in waterproof fabric over his head"*

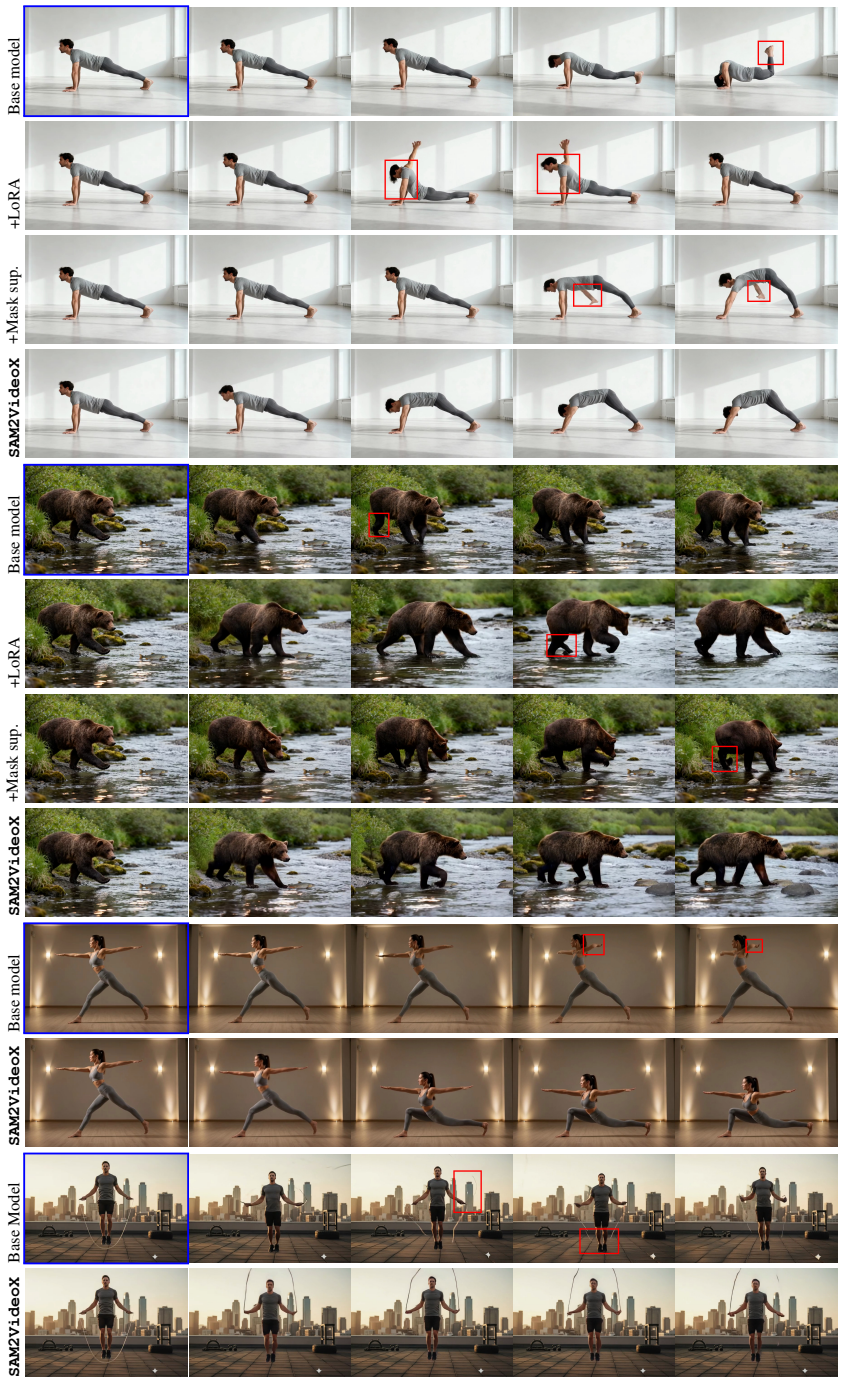

Figure 13: **Extended Visual Comparisons with Baselines.** The blue box denotes the conditioning image. Red boxes highlight structural artifacts in baseline outputs. The first two examples compare the base model, LoRA fine-tuning, mask supervision, and our method (top to bottom), while the last two compare the base model and our method.

4. *"One person is riding on the back of a horse led by another person"*
5. *"A blue fishing boat is navigating in the ocean next to a cruise ship"*
6. *"a brown and white cow eating hay"*
7. *"a couple of boats floating on a body of water"*

8. *"a couple of horses are running in the dirt"*
9. *"a couple of wooden benches in the middle of a street"*
10. *"a dandelion blowing in the wind"*
11. *"a giraffe walking in a field"*
12. *"a group of men riding horses in a dusty arena"*
13. *"a group of motorcyclists racing on a dirt track"*
14. *"a group of people in orange jumpsuits running along a river"*
15. *"a group of people riding bikes down a street"*
16. *"a group of statues on the side of a building"*
17. *"a group of succulents in a rock garden"*
18. *"a highland cow with long horns standing in a field"*
19. *"a large crowd of people walking in a shopping mall"*
20. *"a large rhino grazing in the grass near a bush"*
21. *"a little boy jumping in the air over a puddle of water"*
22. *"a little girl wearing a purple helmet riding a blue bike"*
23. *"a man and a child riding bumper cars in an amusement park"*
24. *"a man and a woman are boxing in a boxing ring"*
25. *"a man and a woman walking under an umbrella next to a brick wall"*
26. *"a man carrying a woman on his back in a field"*
27. *"a man in a mask is walking through a crowd of people"*
28. *"a man in a small boat fishing in the ocean"*
29. *"a man in sunglasses laying on a wooden bench"*
30. *"a man is running on a dirt road"*
31. *"a man on a surfboard riding a wave in the ocean"*
32. *"a man riding a bike down a street"*
33. *"a man riding a horse with a spear in his hand"*
34. *"a man riding a mountain bike on top of a rocky hill"*
35. *"a man sitting on steps playing an acoustic guitar"*
36. *"a man sitting on the ground playing a musical instrument"*
37. *"a man snorkling in the ocean"*
38. *"a man standing on a boat with a net"*
39. *"a man standing on top of a mountain with a backpack"*
40. *"a man standing on top of a sand dune in the desert"*
41. *"a man swinging a tennis racquet at a tennis ball"*
42. *"a panda bear is eating a piece of bamboo"*
43. *"a penguin walking on a beach near the water"*
44. *"a person cooking vegetables in a pan on a stove"*
45. *"a person is putting sauce on a burger"*
46. *"a person jumping in the air over a fence"*
47. *"a person riding a dirt bike"*
48. *"a person riding a horse in a polo match"*
49. *"a person riding a motorcycle down a road"*
50. *"a person riding a skateboard on a concrete floor"*
51. *"a person riding a wave on a surfboard"*
52. *"a person riding an atv on a dirt track"*
53. *"a person slicing salmon on a cutting board"*

54. *"a smiling woman with her hands clasped"*
55. *"two swans swimming on a lake in the fog"*
56. *"a group of hot air balloons flying over a field"*
57. *"a group of people standing on top of a green hill"*
58. *"an aerial view of a group of islands in the middle of a lake"*
59. *"a group of hot air balloons flying over a valley"*
60. *"a room filled with books and teddy bears"*
61. *"a soccer player in a yellow and black shirt is chasing a soccer ball"*
62. *"a squirrel sitting on the ground eating a piece of bread"*
63. *"a tiger walking on a dirt path in the woods"*
64. *"a tiger walking through a wooded area"*
65. *"a view of a snowy mountain side with many buildings"*
66. *"a warthog is walking in the grass"*
67. *"a white rooster standing in a grassy field"*
68. *"a woman carrying a bundle of plants over their head"*
69. *"a woman in a white dress jumping in the air in a field of pink flowers"*
70. *"a woman is making bread in an oven"*
71. *"a woman stirring food in a pan on the stove"*
72. *"a woman walking down a sidewalk with a bag"*
73. *"a woman with curly hair is drinking a beer"*
74. *"a woman with green hair smiling for the camera"*
75. *"a woman with long black hair is posing for a picture"*
76. *"a yellow and white jellyfish is floating in the ocean"*
77. *"a young boy is jumping in the mud"*
78. *"a zebra walking across a dirt road near a field"*
79. *"an elephant walking through a forest"*
80. *"an older man jogging by the water"*
81. *"an older man playing a saxophone on the street"*
82. *"an older woman and a young girl are knitting together"*
83. *"two people in a canoe on a lake with mountains in the background"*
84. *"two people performing a sword fight in front of a forest"*
85. *"two sheep grazing in the grass next to a wooden bridge"*

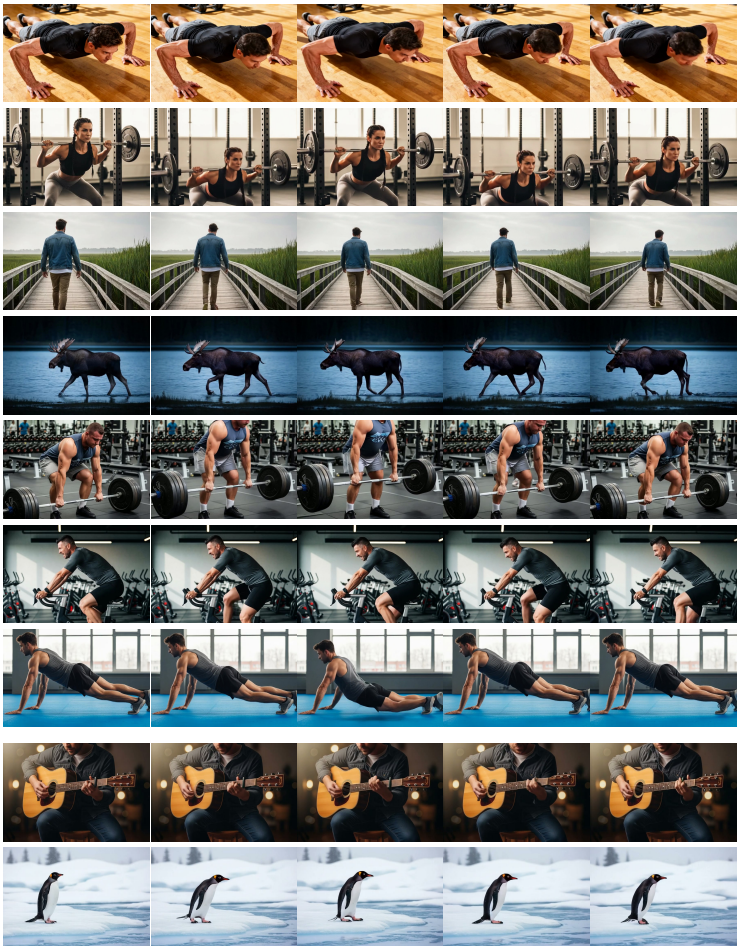

Figure 14: **Additional Qualitative Results.** Representative outputs from `SAM2VideoX`, showcasing structural consistency across diverse motion types.