

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

Anonymous CVPR submission

Paper ID 9200

Abstract

001 *Reality is a dance between rigid constraints and deformable*
002 *structures. For video models, that means generating motion*
003 *that preserves fidelity as well as structure. Despite*
004 *progress in diffusion models, producing realistic structure-*
005 *preserving motion remains challenging, especially for articu-*
006 *lated and deformable objects such as humans and animals.*
007 *Scaling training data alone, so far, has failed to resolve*
008 *physically implausible transitions. Existing approaches rely*
009 *on conditioning with noisy motion representations, such as*
010 *optical flow or skeletons extracted using an external im-*
011 *perfect model. To address these challenges, we introduce*
012 *an algorithm to distill structure-preserving motion priors*
013 *from an autoregressive video tracking model (SAM2) into a*
014 *bidirectional video diffusion model (CogVideoX). With our*
015 *method, we train SAM2VideoX, which contains two innova-*
016 *tions: (1) a bidirectional feature fusion module that ex-*
017 *tracts global structure-preserving motion priors from a re-*
018 *current model like SAM2; (2) a Local Gram Flow loss that*
019 *aligns how local features move together. Experiments on*
020 *VBench and in human studies show that SAM2VideoX de-*
021 *livers consistent gains (+2.60% on VBench, 21–22% lower*
022 *FVD, and 71.4% human preference) over prior baselines.*
023 *Specifically, on VBench, we achieve 95.51%, surpassing*
024 *REPA (92.91%) by 2.60%, and reduce FVD to 360.57,*
025 *a 21.20% and 22.46% improvement over REPA- and LoRA-*
026 *finetuning [11], respectively.*

036 out these constraints, generated motions drift; limbs shear,
037 textures tear, and object identity is lost. Only by achiev-
038 ing this can models move beyond static appearance toward
039 faithful world simulators.

040 Motion remains a major challenge. This is particularly
041 true for articulated and highly deformable objects, such as
042 humans and animals, where models often suffer from inconsis-
043 tent or physically implausible transitions in object states.
044 Inference-time interventions, such as ControlNet-style [37]
045 conditioning on explicit motion representations during in-
046 ference, require knowing the ideal motion *a priori*. Unlike
047 rigid objects whose motion can be well represented by sim-
048 ple dragged trajectories [7], articulated and deformable en-
049 tities lack a unified motion representation. A common as-
050 sumption is that low motion quality stems from insufficient
051 training data, especially for high-quality complex articu-
052 lated motions. However, scaling up or augmenting training
053 data only helps marginally; training with motion proxies
054 (optical flow [4], skeletons [13]) for objects [28, 35] still re-
055 sults in physically implausible transitions. Our experiments
056 show that generated videos still produce lions walking with-
057 out alternating legs and cyclists with static knees. Scaling
058 up such priors results in noisy training data; motion priors
059 are usually collected using imperfect models (*e.g.* RAFT
060 can generate optical flow priors [29]).

061 To improve articulated motion, we propose a simple but
062 powerful idea: deriving *structure from tracking*. Our model,
063 SAM2VideoX, distills *structure-preserving* motion priors
064 from a video tracking model into a video diffusion gen-
065 erator. Previous work has shown that distilling image rep-
066 resentations improves image generation fidelity [38]. To
067 generalize this insight from static to dynamic generation,
068 we distill video representations to improve video genera-
069 tion. Specifically, we leverage SAM2 [23], a state-of-the-art
070 video tracking model trained on large-scale, diverse video
071 data. SAM2 is capable of maintaining object identity across
072 long sequences and through complex occlusions. To track,
073 SAM2’s internal representations have captured how parts
074 move together, how limbs stay connected, and how occlu-
075 sions resolve over time. Instead of conditioning generation

1. Introduction

027 From Heraclitus to Bergson, philosophy has cast cognition
028 as the apprehension of *becoming* rather than *being* [1, 9].
029 While image generation models have excelled at generat-
030 ing what is *there* in high-fidelity images [2, 3, 16], our best
031 video generation still struggles to express the dynamics of
032 *change*. The central challenge is *structure-preserving* motion:
033 dynamics that maintain part topology and local neigh-
034 borhoods while allowing constrained deformations. With-
035



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).

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 ℓ_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 under fast or long-range motion and struggle to maintain object identity through occlusions. Mask-based tracking of-

109
110
111
112
113
114
115
116
117
118

119
120
121
122
123
124
125
126
127
128
129
130
131
132
133

134
135
136
137
138
139

140 fers instance-level signals that are more stable in cluttered
 141 scenes. SAM2 [23] tracks user-prompted regions across
 142 long sequences and is known to generalize well across do-
 143 mains while preserving object identity through occlusions.
 144 However, raw masks are discrete and boundary-focused,
 145 which discards much of the appearance and motion struc-
 146 ture that video diffusion models need. We therefore lever-
 147 age SAM2’s internal features as motion priors: they are
 148 dense, continuous and temporally consistent, and they pro-
 149 vide long-range correspondences that are useful for struc-
 150 ture preservation.

151 **Representation alignment.** Representation alignment was
 152 introduced for image generation by REPA [36] and in-
 153 spired several follow-ups. In the video domain, two main
 154 approaches exist: aligning diffusion features to structured
 155 motion signals such as trajectories or flow [4, 13]; or to
 156 generic video encoders like VideoMAEv2 [30], as in Video-
 157 REPA [38]. Both suffer from key limitations. First, Trajec-
 158 tories and flow provide local supervision and are sensitive
 159 to long-range dynamics, which weakens their usefulness for
 160 structure preservation. Second, popular video encoders like
 161 VideoMAEv2 are optimized for high-level semantic tasks
 162 rather than low-level motion understanding. In contrast, our
 163 approach aligns diffusion features to SAM2’s internal fea-
 164 tures. This prior is unified and structure-centric, allowing
 165 transfer across humans and animals while yielding object-
 166 consistent motion representations without requiring a spe-
 167 cific controller at inference.

168 3. Preliminaries

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

175 **Latent Video Diffusion Model.** Diffusion models gen-
 176 erate samples by inverting a forward noising process [8,
 177 26, 27]. In the latent setting, a pre-trained autoencoder,
 178 with encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$, maps a video $\mathbf{x} =$
 179 $\{I_0, \dots, I_{N-1}\} \in \mathbb{R}^{N \times H \times W \times C}$ to a latent representation
 180 $\mathbf{z} = \mathcal{E}(\mathbf{x})$, where $z \in \mathbb{R}^{N' \times H' \times W' \times C'}$. At timestep t ,
 181 noisy latent \mathbf{z}_t is sampled as

$$182 \mathbf{z}_t = \alpha_t \mathbf{z} + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

183 A DiT f_θ is trained to predict the velocity target $\mathbf{v} = \alpha_t \boldsymbol{\epsilon} -$
 184 $\sigma_t \mathbf{z}$ under the standard v-prediction objective. At inference,
 185 iterative denoising maps $\mathbf{z}_T \rightarrow \mathbf{z}_0$, after which the decoder
 186 $\mathcal{D}(\mathbf{z}_0)$ produces the final video.

187 **Segment-Anything Model 2 (SAM2).** The Segment-
 188 Anything Model 2 (SAM2) extends the image segmentation

189 capability of SAM 1 [15] to the video domain, i.e., tracking
 190 an object in a video conditioned on given prompts such as
 191 points or boxes.

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

$$202 \mathbf{M} = \{M_0, M_1, \dots, M_{N-1}\}$$

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

$$204 \mathbf{F}_{\text{mem}} = \{F_{\text{mem},0}, F_{\text{mem},1}, \dots, F_{\text{mem},N-1}\}.$$

205 4. Method

206 We distill SAM2’s motion structure prior into the video
 207 DiT model by aligning their internal feature representations.
 208 This is achieved through a learnable feature alignment mod-
 209 ule (Sec. 4.1) that projects intermediate DiT features into a
 210 latent space where they can be matched to those of SAM2.
 211 To align their relational motion structures, we introduce a
 212 novel Local Gram Flow (LGF) feature matching operator
 213 (Sec. 4.2). Furthermore, due to the autoregressive nature
 214 of SAM2’s feature, in contrast to the bidirectional features
 215 in DiT, we propose a method that combines SAM2’s for-
 216 ward and backward video features into a single bidirectional
 217 representation, providing a more suitable teacher signal to
 218 be distilled from (Sec. 4.3). Finally, a Local Gram Flow
 219 motion distillation loss (Sec. 4.4) is applied to enforce this
 220 alignment in the latent space. Fig. 2 provides an overview
 221 of our full method.

222 4.1. Feature Alignment Network

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

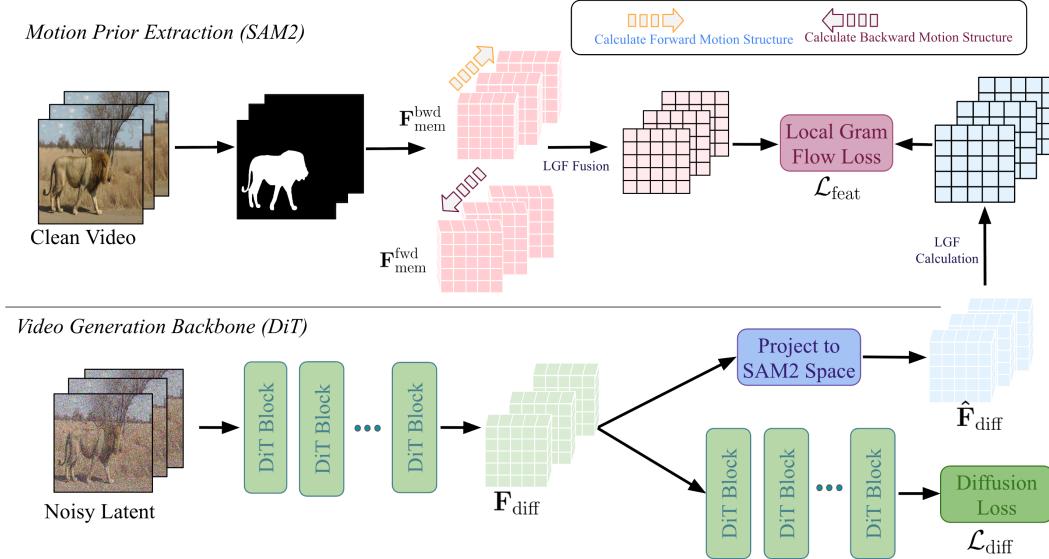


Figure 2. Method overview. The framework consists of two parallel branches: **(Top)** The *Motion Prior Extraction* branch extracts forward and backward memory features ($\mathbf{F}_{\text{mem}}^{\text{fwd}}, \mathbf{F}_{\text{mem}}^{\text{bwd}}$) from SAM2 given a clean video, and fuses them into a bidirectional teacher representation. **(Bottom)** The *Video Generation Backbone* takes noisy latents as input, and the intermediate DiT features \mathbf{F}_{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}_{diff} and \mathbf{F}_{SAM2} , We project \mathbf{F}_{diff} into SAM2's feature space. Specifically, we add a projection module \mathcal{P} on top of \mathbf{F}_{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}_{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}_{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 7×7 spatial neighborhood in the subsequent frame (see Fig. 3). This yields local similarity

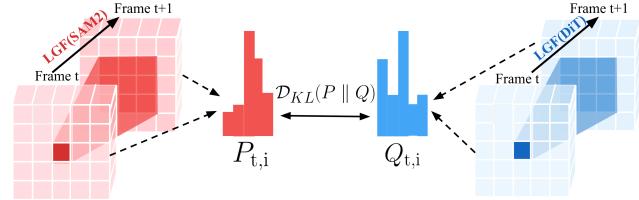


Figure 3. Illustration of the Local Gram Flow (LGF) Loss. The LGF operator captures motion structure by computing similarities between a token at frame t and its 7×7 spatial neighborhood in the subsequent frame $t + 1$. Instead of matching absolute values (e.g., ℓ_2), we convert the resulting similarity vectors ($P_{t,i}, Q_{t,i}$) into probability distributions and align them using the KL Divergence. This forces the model to learn relative motion patterns, not just feature values.

vectors at each position that model likely motion trajectories to 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}_{mem} at each timestep to encode only current and past frames. To bridge this gap and create a teacher feature \mathbf{F}_{SAM2} that is aware of the full video context (similar to

272 DiT), we construct it from both a forward and backward
 273 pass of SAM2. The backward pass is obtained by feeding
 274 the temporally reversed video into SAM2 to extract
 275 its backward features. After obtaining the backward fea-
 276 tures, we remap them to the original temporal order (i.e.,
 277 $t \mapsto N-1-t$) to align with the forward features, produc-
 278 ing:

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

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

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

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

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

295 4.4. Motion Distillation Loss

296 Finally, the motion distillation loss, $\mathcal{L}_{\text{feat}}$, is designed to
 297 match the LGF distributions of the student $\hat{\mathbf{F}}_{\text{diff}}$ and the
 298 fused teacher \mathbf{F}_{SAM2} , rather than enforcing one-to-one cor-
 299 respondence as in a standard ℓ_2 loss. More specifically, we
 300 align the distribution of spatio-temporal motion similarities
 301 between the video DiT feature and the SAM2 feature. We
 302 therefore apply a softmax to each token’s similarity vector
 303 (turning it into a probability distribution) and measure the
 304 distance using the KL divergence. This approach focuses
 305 on the relative ranking of similarities, which we argue bet-
 306 ter captures the underlying motion structure. Specifically,
 307 we compute probability distributions:

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

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

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

$$312 \mathcal{L}_{\text{feat}}\left(\hat{\mathbf{F}}_{\text{diff}}, \mathbf{F}_{\text{SAM2}}\right) = \frac{1}{(N'-1)|\Omega|} \sum_{t=0}^{N'-2} \sum_{i \in \Omega} \text{KL}(P_{t,i} \| Q_{t,i})$$

313 Our ablation studies (Table 2) empirically validate that
 314 this LGF-KL combination is critical, yielding significant
 315 gains over simpler alternatives (e.g., LGF with ℓ_2 loss, or
 316 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_{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 pre-compute 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 1×10^{-4} and momentum parameter $(\beta_1, \beta_2) = (0.9, 0.95)$. We train for 3,000 steps on $8 \times \text{H200}$ 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

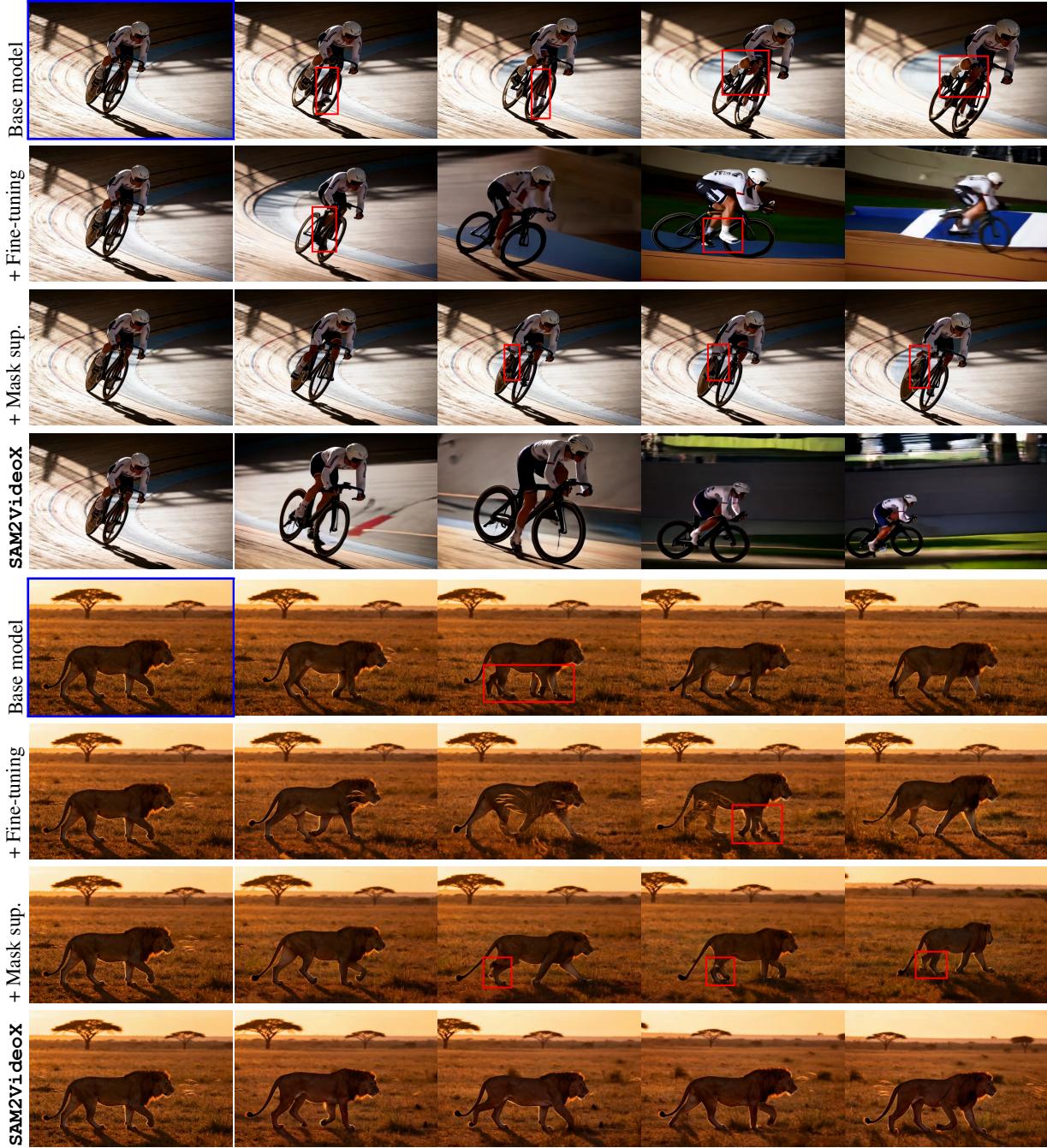


Figure 4. **Qualitative comparison on articulated motion.** The blue box indicates the input image. Red boxes highlight common failure modes in baselines, including structural distortion (cyclist’s legs) and physical implausibility (three-legged lion, inconsistent leg motion). In contrast, our method (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.

369 conduct a double-blind user study using 40 randomly sam-
 370 pled prompts. Participants are presented with two side-by-
 371 side videos (Ours vs. Baseline) in random order and asked
 372 to select the preferred one based on motion smoothness and
 373 subject consistency. All methods generate 49-frame videos

374 at 8 fps and 720×480 resolution, using 50 denoising steps
 375 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

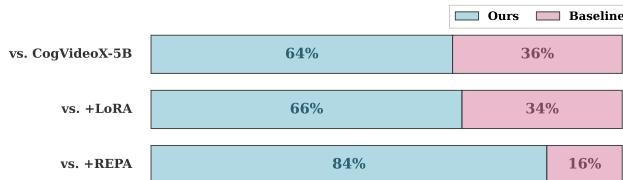


Figure 5. Human preference win rates. Participants were shown pairwise comparisons and asked to select the video with ‘superior limb consistency and fewer motion artifacts’. Our method was strongly preferred ($> 64\%$) over all baselines, confirming its superior ability to generate plausible, artifact-free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

376

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.

399
400
401
402
403
404
405

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.

406
407
408
409
410
411
412
413
414
415
416

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.

417
418
419
420
421
422
423
424
425
426
427
428

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.

430
431
432
433

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 ℓ_2 loss ("w/o LGF") and an ℓ_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 ℓ_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 (ℓ_2).

434

LGF fusion is essential to resolve bidirectional conflicts. We validate our teacher design in Table 2. While using only the forward stream ($\mathbf{F}_{\text{mem}}^{\text{fwd}}$) 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.

449

KL divergence outperforms ℓ_2 for structural alignment. We further ablate the loss function design given our LGF teacher. As shown in Table 2, applying a standard ℓ_2 loss

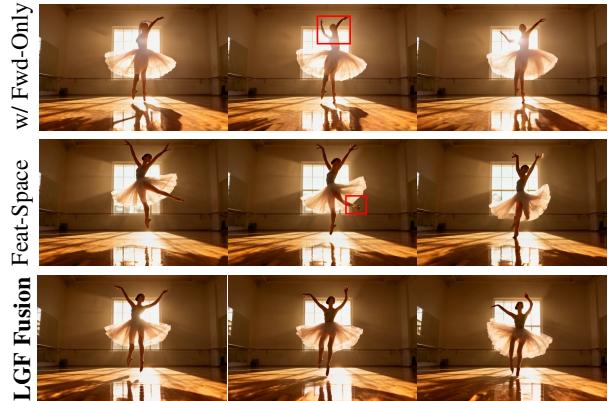


Figure 7. **Qualitative ablation of our bidirectional LGF fusion.** (Top) The forward-only teacher ("w/ Fwd-Only") produces artifacts, like the ballerina's arm 'folding' incorrectly (red box). (Middle) "Feat-Space Fusion" causes structural tearing. (Bottom) Our "LGF Fusion" correctly preserves the limb's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

452
453
454
455
456
457
458
459
460
461
462
463
464
465466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481

482 **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, Jia-hao 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. 1, 2
- [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
- [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
- [14] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker3: 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
- [17] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shanghai Yuan, Lihuahan 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, Jiaoqiao 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-*

- 595 *national conference on computer vision*, pages 4195–4205,
596 2023. 2, 3
- 597 [21] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young,
598 Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu,
599 Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang
600 Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu,
601 Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei,
602 Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang,
603 Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao,
604 Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training
605 a commercial-level video generation model in \$200k.
606 *arXiv preprint arXiv:2503.09642*, 2025. 2
- 607 [22] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc
608 Van Gool, Markus Gross, and Alexander Sorkine-Hornung.
609 A benchmark dataset and evaluation methodology for video
610 object segmentation. In *Proceedings of the IEEE conference*
611 *on computer vision and pattern recognition*, pages 724–732,
612 2016. 2
- 613 [23] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang
614 Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman
615 Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2:
616 Segment anything in images and videos. *arXiv preprint*
617 *arXiv:2408.00714*, 2024. 1, 3
- 618 [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-
619 net: Convolutional networks for biomedical image segmen-
620 tation. In *Medical image computing and computer-assisted*
621 *intervention-MICCAI 2015: 18th international conference,*
622 *Munich, Germany, October 5-9, 2015, proceedings, part III*
623 18, pages 234–241. Springer, 2015. 2
- 624 [25] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico
625 Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov,
626 Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa,
627 Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan
628 Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana,
629 Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt,
630 Camille Couprie, Julien Mairal, Hervé Jégou, Patrick La-
631 batut, and Piotr Bojanowski. DINOV3, 2025. 5
- 632 [26] Jiaming Song, Chenlin Meng, and Stefano Ermon.
633 Denoising diffusion implicit models. *arXiv preprint*
634 *arXiv:2010.02502*, 2020. 3
- 635 [27] Yang Song and Stefano Ermon. Generative modeling by esti-
636 mating gradients of the data distribution. *Advances in neural*
637 *information processing systems*, 32, 2019. 3
- 638 [28] Keqiang Sun, Dor Litvak, Yunzhi Zhang, Hongsheng Li, Ji-
639 ajun Wu, and Shangzhe Wu. Ponymation: Learning artic-
640 ulated 3d animal motions from unlabeled online videos. In
641 *European Conference on Computer Vision*, pages 100–119.
642 Springer, 2024. 1
- 643 [29] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field
644 transforms for optical flow. In *Computer Vision–ECCV*
645 *2020: 16th European Conference, Glasgow, UK, August 23–*
646 *28, 2020, Proceedings, Part II 16*, pages 402–419. Springer,
647 2020. 1, 2
- 648 [30] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang.
649 Videomae: Masked autoencoders are data-efficient learners
650 for self-supervised video pre-training. *Advances in neural*
651 *information processing systems*, 35:10078–10093, 2022. 3
- [31] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach,
652 Raphael Marinier, Marcin Michalski, and Sylvain Gelly. To-
653 wards accurate generative models of video: A new metric &
654 challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5
- [32] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao,
655 Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianx-
656 iao Yang, et al. Wan: Open and advanced large-scale video
657 generative models. *arXiv preprint arXiv:2503.20314*, 2025.
658 2
- [33] Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng
659 Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun
660 Zhou, Xinze Chen, Guan Huang, et al. Humandreamer:
661 Generating controllable human-motion videos via decoupled
662 generation. In *Proceedings of the Computer Vision and Pat-
663 tern Recognition Conference*, pages 12391–12401, 2025. 5
- [34] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu
664 Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiao-
665 han Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video
666 diffusion models with an expert transformer. *arXiv preprint*
667 *arXiv:2408.06072*, 2024. 2, 5
- [35] Zhangsihao Yang, Mingyuan Zhou, Mengyi Shan, Bingbing
668 Wen, Ziwei Xuan, Mitch Hill, Junjie Bai, Guo-Jun Qi, and
669 Yalin Wang. Omnimotiongpt: animal motion generation
670 with limited data. In *Proceedings of the IEEE/CVF Con-
671 ference on Computer Vision and Pattern Recognition*, pages
672 1249–1259, 2024. 1
- [36] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon
673 Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie.
674 Representation alignment for generation: Training diffu-
675 sion transformers is easier than you think. *arXiv preprint*
676 *arXiv:2410.06940*, 2024. 2, 3, 5, 7
- [37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding
677 conditional control to text-to-image diffusion models. In
678 *Proceedings of the IEEE/CVF international conference on*
679 *computer vision*, pages 3836–3847, 2023. 1, 2
- [38] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing
680 Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Vide-
681 orepas: Learning physics for video generation through re-
682 lational alignment with foundation models. *arXiv preprint*
683 *arXiv:2505.23656*, 2025. 1, 2, 3