arXiv:2609.04383v1 [cs.CV] 3 Sep 2026

What Moves? Localized Motion Representations for Compositional Scene Control

Frank Fundel[*], **Malek Ben Alaya**[*], **Thomas Ressler-Antal**[*]
Stefan Andreas Baumann, **Björn Ommer**
CompVis @ LMU Munich, Munich Center for Machine Learning (MCML)

Real-world dynamics are inherently compositional: multiple entities move simultaneously within a shared scene, each exhibiting distinct motion patterns. Yet most existing video representations encode motion globally, without explicitly capturing localized motion for individual entities. Crucially, motion is defined relative to a global reference frame, including camera motion and scene layout. However, localized embeddings are often computed from cropped images or obtained by masking features after encoding, discarding the context needed to interpret motion. To address this, we introduce a promptable localized motion representation that produces persistent embeddings for user-specified regions defined by spatial masks. Rather than cropping the input or masking features, our model processes the full video and conditions motion encoding directly on the queried region. This yields temporally consistent, region-addressable embeddings that isolate local dynamics while retaining the global context required for disambiguation. We demonstrate object-level motion transfer, enabling controlled composition of dynamic scenes. Beyond generative control, our embeddings support localized action classification in multi-actor videos. Across both tasks, our approach improves controllability and outperforms global representations localized through cropping or post-hoc masking.

Project Page: https://compvis.github.io/WhatMoves
Code: https://github.com/CompVis/WhatMoves

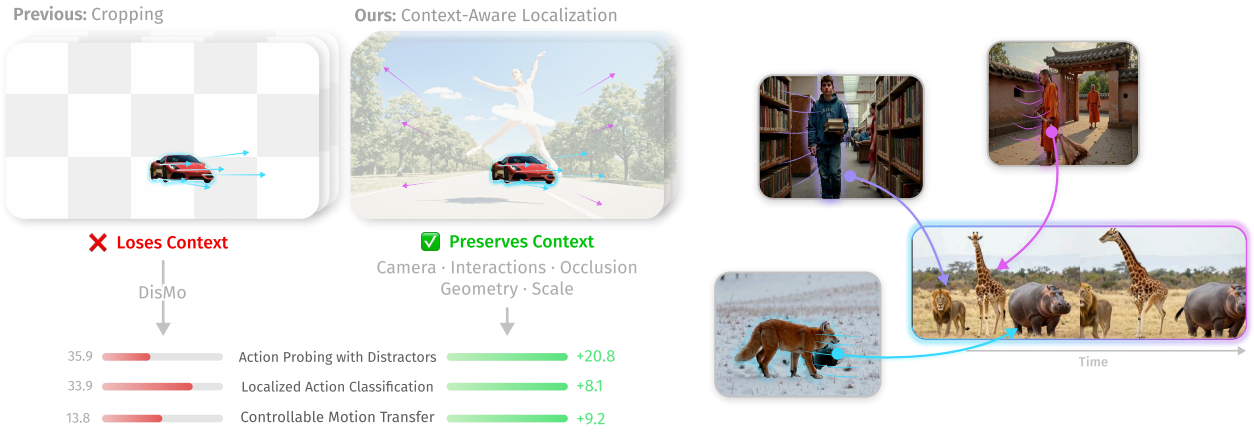


Figure 1: Localized motion representations enable context-aware scene composition. (Left) Our method extracts motion from selected regions while preserving global contextual cues such as camera motion, interactions, and scale, unlike cropping-based localization. **(Right)** These localized representations can be transferred from multiple source videos to target objects, enabling the composition of distinct motions within a single scene.

1 Introduction

Motion is a defining characteristic of the visual world. Real-world scenes are inherently compositional, consisting of multiple entities whose dynamics unfold simultaneously within a shared environment. A driving scene contains vehicles,

[*]Equal contribution.

pedestrians, and background motion with distinct temporal patterns; human-object interactions involve coordinated, yet separable movements of body parts and manipulated objects. Modeling such dynamics requires representations that are localized, persistent, and semantically grounded. Despite rapid advances in video generation and motion control, most existing approaches represent motion either globally or at the pixel level. Global motion embeddings entangle the dynamics of all entities into a single latent code [39], while dense optical flow or trajectory fields encode motion locally but lack semantic structure and compositional interpretability. As a result, motion remains difficult to isolate, recombine, or selectively manipulate at the object level. This limitation becomes particularly apparent in tasks such as motion transfer and video editing. Existing methods often extract motion globally and reapply it to new scenes [13, 32, 39, 41–43], leading to motion leakage, spatial ambiguity, or incorrect attachment to unintended regions. While these approaches demonstrate impressive generative quality, they lack an explicit, semantic representation of who is moving and how. A seemingly straightforward solution is to localize motion by applying global motion encoders to spatial crops of a video. Yet motion is inherently relational: its meaning depends on global reference frames and interactions between entities. When this context is removed, the motion signal within a crop may no longer uniquely determine the underlying dynamics. Consequently, embeddings extracted from isolated regions are not guaranteed to represent well-defined object-level motion.

Our approach operates directly on the full input video while leveraging spatiotemporal spatial masks to localize motion signals. Instead of cropping regions, we retain the entire scene context and apply region-specific masks over time within the encoder. A motion encoder processes the masked video features to produce compact entity-level motion embeddings, while a motion decoder reconstructs the corresponding localized point trajectories, encouraging spatial specificity and disentanglement from static appearance. By preserving global context during encoding and enforcing localization through masking, our method yields motion representations that are both contextually grounded and spatially isolated. This structured representation naturally enables compositional video generation and control. Motion from individual entities can be selectively extracted, recombined, and reattached to arbitrary targets to animate a scene. Beyond motion transfer, the learned embeddings support localized action classification, highlighting their role as general representations for motion understanding and generation.

Through extensive experiments, we demonstrate that our approach enables precise spatial controllability and structured motion composition in generative settings. In localized motion transfer experiments, our method supports accurate extraction and reattachment of region-specific dynamics while significantly reducing motion leakage compared to global motion conditioning. It further enables compositional recombination of motions across scenes and more consistent cross-view transfer than track-conditioned models that rely on viewpoint-specific trajectories. Beyond generation, we evaluate the learned motion embeddings on motion understanding tasks. In localized action classification, our representations capture spatially specific motion patterns that generalize across actors and scenes, outperforming global video embeddings localized via cropping or post-hoc masking. Together, these results highlight the importance of encoding local motion while preserving full-scene context for both controllable generation and motion understanding.

Our main contributions are summarized as follows:

- We introduce a promptable localized motion representation that encodes persistent, region-specific motion embeddings conditioned on spatial regions.
- We show that encoding local motion requires full-scene context, and demonstrate through localized action classification that crop-based or post-hoc masking approaches are insufficient.
- We demonstrate that the learned motion embeddings enable object-level motion transfer as well as selective extraction and recombination of region-level dynamics, supporting controllable and compositional scene animation in image-to-video settings.

2 Related Work

Video Generation Models. Diffusion-based generative models have driven rapid progress in video synthesis. Building on advances in text-to-image generation [16, 40, 45], many works extend pretrained image backbones with temporal modules to generate coherent video sequences [6, 14]. Scaling model capacity and training data has further improved temporal consistency and semantic richness in text-to-video (T2V) and image-to-video (I2V) systems [3, 4, 15, 24, 52, 62]. Despite these advances, most video generation systems control motion primarily through text prompts, which provide

only coarse and indirect guidance over how objects move. We introduce an object-centric representation of motion that enables targeted manipulation of dynamics at the level of individual entities. Our approach can be integrated with existing video generation models, enabling more precise and structured control over motion without redesigning the underlying generative architecture.

Motion Control in Video Generation. Recent work improves controllability in video diffusion models by introducing explicit motion guidance signals. A common approach relies on image-plane geometric control, where motion is specified through trajectories, bounding boxes, or structural signals such as optical flow and depth [9, 21, 53–55, 65, 66]. Subsequent works extend this paradigm with trajectory-aware conditioning mechanisms and motion-guided attention modules that enable more precise point-level motion control and multi-entity interactions [1, 11, 44, 56, 67, 68]. Other approaches guide generation directly using motion signals such as optical flow, temporally warped noise, or related geometric cues [5, 25, 71]. While these approaches provide strong geometric control when the guidance signal aligns with the target layout, motion is typically represented in pixel space through trajectories, flow fields, or related spatial signals. As a result, the encoded dynamics remain tightly coupled to viewpoint, object layout, and scene configuration, making it difficult to transfer the same motion across changes in camera perspective, object shape, or scene structure.

A complementary line of work studies motion representation learning and motion transfer, where dynamics are extracted from a source video and applied to a different scene. Several diffusion-based text-to-video approaches leverage internal features of pretrained video diffusion models to encode and propagate motion across scenes [22, 28, 36, 42, 57, 64, 69]. More recent work explores motion-guided editing paradigms that modify subject appearance while preserving source motion [29, 30]. Other approaches focus on efficiency or representation abstraction, such as training-free motion transfer formulations or depth-aware motion embeddings [32, 41]. Methods operating on lower-level features often inherit geometric constraints from the source video, while higher-level feature-based approaches provide greater flexibility but may entangle motion with appearance, leading to content leakage. Recent work proposes more abstract motion embeddings to mitigate this entanglement [20, 39]. However, these approaches typically represent motion globally or as latent feature dynamics within a generative backbone, without explicit object-level structure or spatial localization. Consequently, they offer limited control over where motion is extracted from and how it is distributed across entities in multi-subject scenes.

Our work differs in that we learn promptable, localized motion representations conditioned on spatial regions. Unlike trajectory-based control methods that rely on image-plane guidance signals, and motion transfer approaches that represent dynamics globally within a generative backbone, our approach models motion at the object level while preserving full-scene context.

3 Method

Modeling video dynamics in a structured and controllable manner requires motion representations that are both spatially localized and grounded in global scene context. While global motion embeddings capture overall dynamics, they do not provide entity-level motion units that can be independently extracted and recombined. In this section, we first examine the challenges of localizing motion in video and motivate a context-aware formulation. We then introduce our promptable localized motion representation.

3.1 Motivation: Localized Motion Requires Global Context

A natural strategy for isolating entity motion is to encode motion from spatial crops. However, unlike appearance, motion is inherently *relational*: its interpretation depends on global context, camera motion, and interactions between objects. For instance, an object moving away from the camera may produce the same local motion pattern as a camera zoom-out. Similarly, object motion and camera motion often become indistinguishable when observed in isolation. These ambiguities arise because spatial cropping removes the reference frame needed to interpret motion. Consequently, crop-based motion encoders frequently learn ambiguous dynamics or collapse to dominant camera motion. Reliable entity-level motion representations therefore require localization *without discarding global context*. This observation motivates our formulation: the encoder processes the *entire video*, while entity-specific motion is extracted through structured prompting and masked supervision.

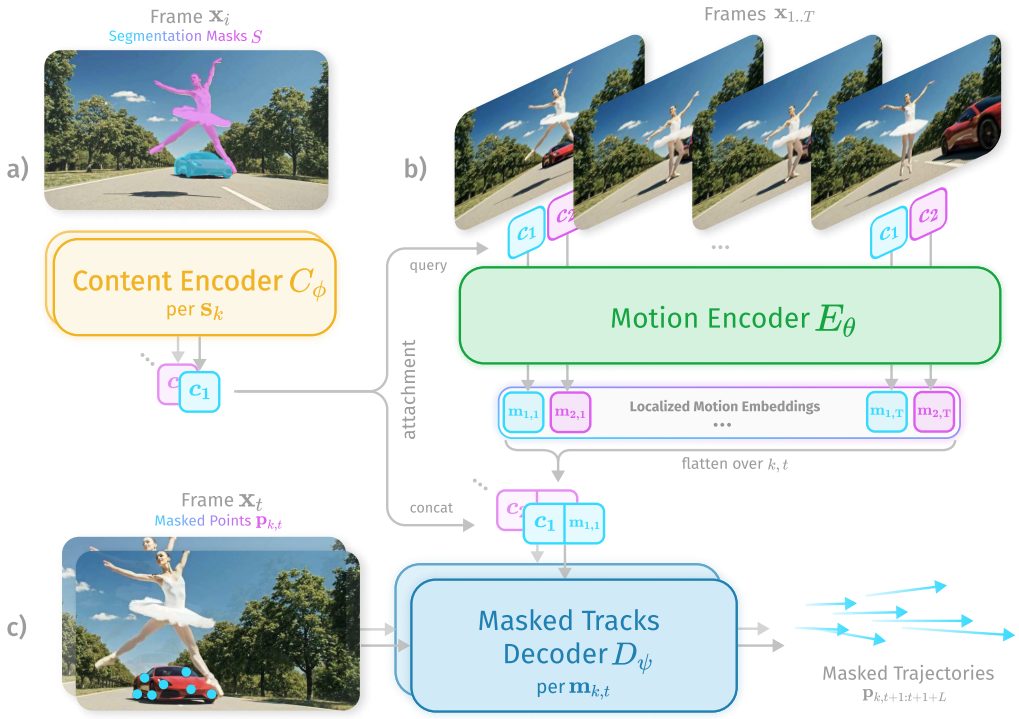


Figure 2: Method Overview. Given a video frame with corresponding object masks, the content encoder produces static region tokens for appearance grounding **(a)**. The motion encoder **(b)** extracts persistent, region-specific motion tokens while preserving global context. The masked flow decoder **(c)** predicts region-level point trajectories, enforcing localized motion supervision within masked areas.

3.2 Promptable Localized Motion Embeddings

Building on this observation, we introduce a promptable localized motion representation that processes the full video while extracting region-specific motion embeddings conditioned on spatiotemporal masks. Each embedding captures the temporal dynamics of a queried entity and remains persistent over time. By preserving global context during encoding and enforcing region-level aggregation through masking, our framework produces motion representations that are both spatially precise and contextually grounded.

Consider a video $V = \{\mathbf{x}_t\}_{t=1}^{T}$ with T frames. In the first frame $\mathbf{x}_1$, we are given region masks $S := \{\mathbf{s}_k\}_{k=1}^{K}$. Our goal is to extract persistent, region-specific motion tokens $\mathbf{m}_k := \{\mathbf{m}_{k,t}\}_{t=1}^{T}$, where $\mathbf{m}_{k,t}$ describes the motion of region k at time t. We tackle this problem using three components: (1) a **motion encoder** that extracts localized motion tokens from the full video, (2) a **content encoder** that provides appearance grounding for the queried entity, and (3) a **masked tracks decoder** that supervises motion tokens through future trajectory prediction. An overview of our framework is depicted in Fig. 2.

Motion Encoder. The motion encoder E_θ maps the full video V and a set of entity content embeddings $\mathbf{c} = \{\mathbf{c}_k\}_{k=1}^{K}$ to temporally evolving motion tokens $\mathbf{m} = \{\mathbf{m}_k\}_{k=1}^{K}$, i.e., $E_\theta : (V, \mathbf{c}) \mapsto \mathbf{m}$. Importantly, E_θ observes the entire scene, allowing each entity-specific motion token to use global context such as camera motion, occlusions, and object interactions. The content embeddings serve as region-specific queries, encouraging $\mathbf{m}_{k,t}$ to capture the motion of the corresponding entity while reducing leakage from other regions.

Content Encoder. The content encoder C_ϕ provides appearance grounding by mapping a frame $\mathbf{x}_i$ and region mask $\mathbf{s}_k$ to a content token $\mathbf{c}_k$, *i.e.*, $C_\phi : (\mathbf{x}_i, \mathbf{s}_k) \to \mathbf{c}_k$. This token captures the semantic identity of the prompted region and is time-invariant. It queries the motion encoder to specify the entity whose motion should be extracted and is passed to the tracks decoder to reattach motion embeddings to the correct region, discouraging them from encoding spatial identity.

Masked Tracks Decoder and Objective. We supervise the motion tokens with a masked tracks decoder D_ψ. For each entity k, sparse query points are sampled only inside its region S_k and tracked through the video, yielding trajectories $\mathbf{p}_{k,1:T}$. At a valid timestep t, the decoder receives the current frame $\mathbf{x}_t$, query positions $\mathbf{p}_{k,t}$, the content embedding

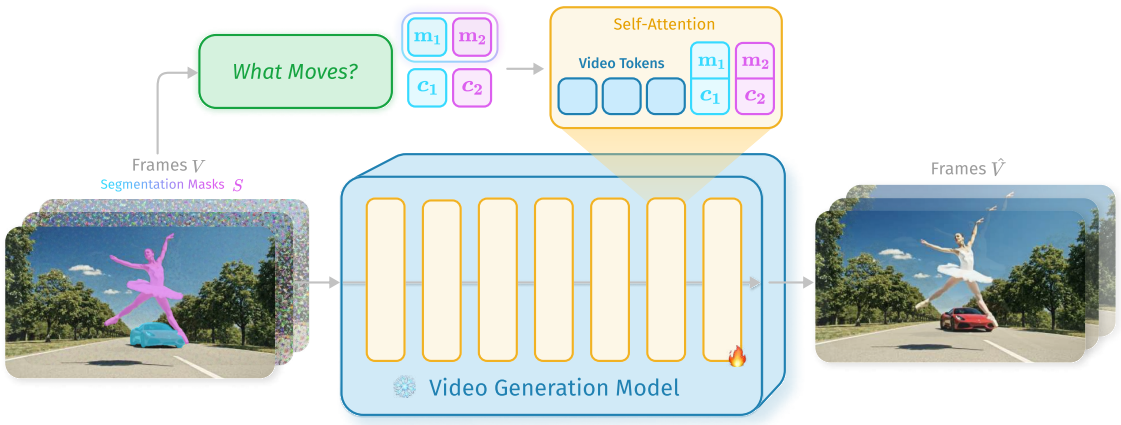


Figure 3: Video Model Conditioning. With our localized representations, we fine-tune a large video generation model. For this, we append to each temporal self-attention layer of the the model, the respective motion embeddings $\mathbf{m}_{1:K,t}$ and content embeddings $\mathbf{c}_{1:K}$ for all regions K.

$\mathbf{c}_k$, and the motion token $\mathbf{m}_{k,t}$, and predicts future track positions $\mathbf{p}_{k,t+1:t+L}$. Although this conditional trajectory distribution could be modeled generatively, e.g. with a flow matching decoder, we use a deterministic decoder trained with an ℓ_2 reconstruction loss:

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{V,k,t}\left[\|D_\psi(\mathbf{x}_t, \mathbf{p}_{k,t}, \mathbf{c}_k, \mathbf{m}_{k,t}) - \mathbf{p}_{k,t+1:t+L}\|_2^2\right],$$

evaluated only for valid future track positions within a region, encouraging each motion token to encode localized motion rather than unrelated scene dynamics.

3.3 Localized Motion-Guided Video Generation

Once trained, the motion encoder extracts region-specific motion tokens from a source video by querying a spatial region. These tokens capture the dynamics of the selected entity over time and enable localized motion transfer, where motion from a source region is attached to a specified object in a target scene. To generate videos following these dynamics, we condition a pretrained video diffusion model on sequences of localized motion tokens. For each region k and timestep t, we concatenate the motion tokens m_k, t with their corresponding content token c_k,

$$\mathbf{q}_{k,t} = [\mathbf{m}_{k,t}; \mathbf{c}_k], \tag{1}$$

and append the resulting conditioning tokens to the video token sequence $\mathbf{V}_t = \{\mathbf{v}_{t,n}\}_{n=1}^N$ processed by the model's self-attention layers:

$$\mathbf{V}'_t = [\mathbf{V}_t, \mathbf{q}_{1..K,t}]. \tag{2}$$

The backbone is fine-tuned using LoRA [18] to incorporate the additional conditioning while preserving the pretrained weights. This fine-tuning approach is illustrated in Fig. 3.

4 Experiments

To evaluate the properties of our motion representations, we conduct experiments across several settings. We study controllable motion transfer to assess whether motion can be transferred across entities and scenes enabling compositional scene control, localized action classification to test whether motion can be extracted for a queried subject in the presence of distractors, and global motion understanding tasks to evaluate whether the representations retain semantic motion properties beyond localized queries.

Implementation Details. We implement our motion encoder as a 3D Vision Transformer [8, 86M parameters], preceded by a DINOv2-B [35] frame embedding stage. The masked flow decoder is also implemented as a Transformer[86M

parameters]. We jointly train the encoder and decoder using AdamW [31] for 600k steps with a batch size of 128 on open-world videos from OpenVid-1M [34] and an internally collected dataset, enabling the model to learn general-purpose motion representations. For motion transfer tasks, we utilize CogVideoX-5B [63] as a backbone. We provide further implementation details in the supplementary material.

4.1 Motion Transfer

Localized Motion Attachment. In this experiment, we evaluate whether our framework can attach motion from a source region to a specified target region. To this end, we construct a benchmark that tests whether motion transfer methods can control *which* subject receives transferred motion. Each source video mostly contains a single subject, while the target initial frame always contains multiple subjects. Our method transfers motion from the source subject to a designated target subject, whereas baseline methods without subject-level prompting perform global motion transfer without explicit target selection. This setting enables direct evaluation of whether motion can be accurately directed to a specified subject while suppressing unintended propagation to others. We select source videos from the MTBench dataset [42], originally gathered from DAVIS [37] and VOS [60], and extract subject masks using the promptable segmentation model SAM2 [38]. For each source video, we generate multiple target initial frames using Z-Image Turbo [49].

We quantitatively compare methods using two established global metrics: *Temporal Consistency*, measured via alignment of CLIP embeddings between consecutive generated frames; and *Driving Video Similarity*, computed as the alignment between CLIP embeddings of source and generated video frames. To specifically assess subject-level controllability, we compute an *in-region fidelity* score using a tracking-based Chamfer distance between the source subject trajectories and the trajectories of the designated target subject in the generated video. Higher values indicate more accurate transfer of motion to the intended subject. To quantify unintended motion propagation, we additionally compute an *out-of-region leakage* score using the same Chamfer-based metric between the source subject trajectories and all non-target subject trajectories in the generated video. Lower values indicate better suppression of motion leakage.

We compare against two classes of motion transfer methods: low-level approaches, which provide localized control but rely on low-level motion signals such as tracks, and semantic motion transfer methods, which utilize higher-level motion representations but operate globally. Naturally, low-level methods often achieve strong scores on motion fidelity metrics such as in-region motion fidelity, since they directly follow the source trajectories. However, these approaches do not generalize well across viewpoint changes or structural differences between source and target subjects. In contrast, semantic methods capture higher-level motion patterns but typically lack spatial control. This leads to higher motion leakage, as motion is transferred to arbitrary subjects. We illustrate these trade-offs qualitatively in Fig. 4. Our approach combines both properties by enabling localized attachment of semantic motion. As shown in Tab. 1, our method achieves strong in-region motion fidelity while substantially reducing motion leakage compared to global semantic conditioning. Although ATI [1] achieves the highest fidelity, it exhibits significantly higher out-of-region leakage. Compared to the semantic baseline DisMo [39], our method improves in-region motion fidelity while improving localization. Overall, our approach achieves a favorable trade-off between motion fidelity and spatial isolation. Since our representation captures semantic motion rather than raw trajectories while remaining spatially localized, it generalizes to transfers across viewpoint changes and structural differences between source and target subjects.

Composing Scenes. Beyond transferring motion to a single subject, compositional scene generation requires the ability to combine motion from multiple sources and assign them to different subjects within the same scene. To evaluate this capability, we extend our evaluation to a *multi-source* setting where motion is extracted from multiple source videos and attached to different target subjects in a shared target frame. Each source video provides motion for a distinct subject, and our method is prompted to compose these motions within a single generated scene.

As shown in Tab. 3, our method shows a competitive in-region motion fidelity, while substantially reducing out-of-region motion leakage compared to all baselines. This indicates that our localized motion representations enable controlled composition of multiple motion sources within the same scene, allowing motion to be selectively attached to the intended subjects without affecting others. Qualitative examples are shown in the supplementary material.

We additionally conduct a blinded user study using non-cherry-picked examples, with the full experimental setup

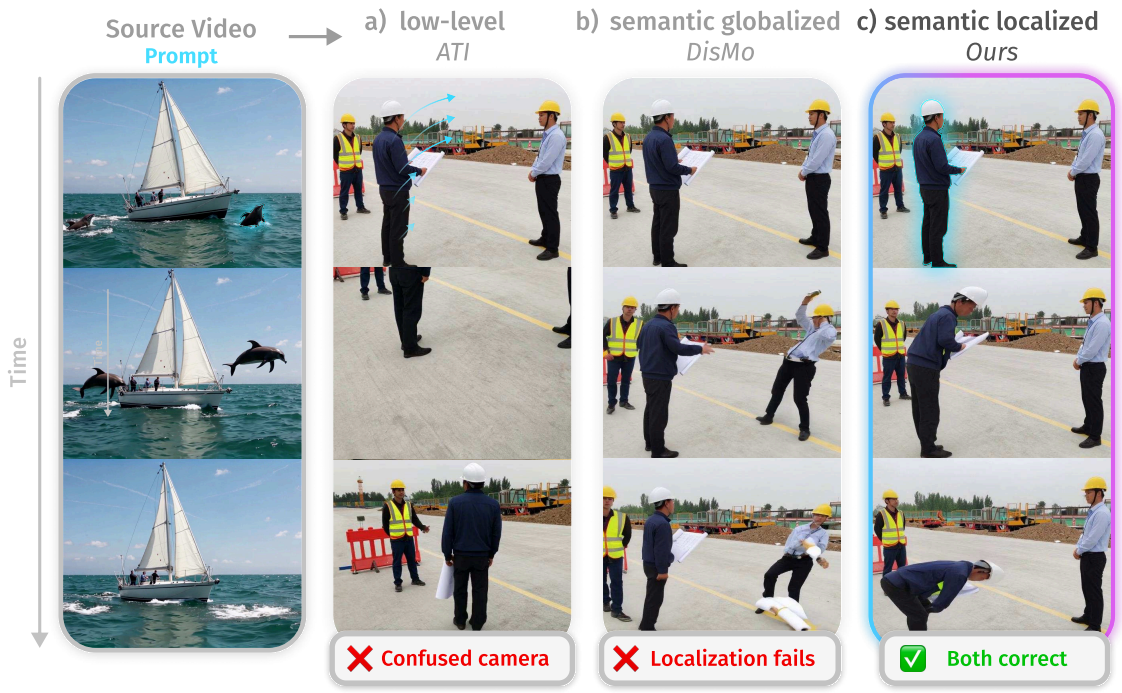


Figure 4: Motion Transfer Paradigms. Existing methods are typically either localized or semantic. **(a)** Low-level approaches transfer exact motion and implicitly assume identical structure; while localized, they do not generalize across different structures. **(b)** Semantic motion transfer methods generalize motion across structures but operate globally and lack localization. **(c)** Our approach enables motion transfer that is both semantic and localized, bridging the two paradigms.

provided in the supplementary material. Participants independently assess each generated video along three dimensions: overall visual realism, the quality and spatial locality of the transferred motion, and the presence of visible artifacts. Tab. 2 shows that across the evaluated methods, our approach is consistently preferred for both realism and motion transfer quality. In particular, it outperforms the next-best baseline by at least 27.78% in realism and 9.8% in motion transfer quality, indicating that the generated videos are perceived as both more visually plausible and more effective at transferring motion to the intended target region.

4.2 Localized Action Classification

A key property of localized motion representations is the ability to capture and semantically describe the motion of a subject of interest while remaining invariant to its appearance and unaffected by other entities in the scene. To evaluate this capability, we test whether motion embeddings extracted from a spatial query encode the underlying action dynamics independently of visual appearance and the actions performed by other actors.

Evaluation Setup. We consider a *zero-shot localized action classification* task: given a video and a spatial query specifying an actor, the goal is to classify the action performed by that actor using only the extracted motion embeddings. We construct this benchmark using the A2D dataset [58], which consists of videos depicting multiple actors, each annotated with segmentation masks and an associated action label. For each actor–action instance, we extract a motion embedding conditioned on the corresponding segmentation mask and classify the action using a k-nearest neighbor (kNN) classifier. To ensure the evaluation emphasizes motion rather than appearance, instances of the same actor are excluded from the retrieval pool. Correct classification therefore requires identifying the underlying motion pattern rather than relying on actor identity or visual cues. We compare against image-, video-, and motion-based methods and report top-1 Accuracy and weighted F1 score to account for class imbalance. To adapt the other methods to this localized setting, we either crop the input video (RGB), or the resulting spatial latent representations (LATENT). If applicable, we report results for both strategies. We additionally construct and evaluate two constrained versions of our method, one where we apply the aforementioned cropping mechanism (**Ours**$_{\text{RGB}}$), and one where we do not apply localized querying at all (**Ours**$_{\text{GLOBAL}}$). Results are summarized in Tab. 4.

Table 1: Controllable Motion Transfer. We report temporal consistency, source similarity, in-region motion fidelity, out-of-region motion leakage, and motion selectivity, defined as the difference between fidelity and leakage. Our method outperforms semantic motion transfer approaches and achieves the best overall motion selectivity, demonstrating the strongest balance between accurately transferring motion to the target region and preventing leakage to the rest of the scene.

		video model properties		motion transfer properties		
	Model	**Temporal Consistency** ↑	**Source Similarity** ↑	**In-Region Motion Fidelity** ↑	**Out-of-Region Motion Leakage** ↓	**Motion Selectivity** ↑
low-level	ATI [1] (Wan 2.1-14B [52])	0.9725	0.5930	**0.8111**	0.6108	0.2003
	WanMove [7] (Wan 2.1-14B)	0.9738	0.5949	0.7438	0.5608	0.1830
	Tora [67] (CogVideoX-5B [63])	**0.9758**	0.5807	0.7021	**0.4738**	0.2283
semantic	DisMo [39] (CogVideoX-5B)	0.9657	**0.5947**	0.6981	0.5604	0.1377
	Ours (CogVideoX-5B)	**0.9690**	0.5888	**0.7462**	**0.5158**	**0.2304**



Figure 5: Composite Qualitative Examples. We provide several examples, showcasing our capability to compose scenes from localized motion embeddings of multiple source videos. Our model correctly isolates and attaches the motion of multiple subjects into one coherent scene.

Context is Important for Localized Motion Understanding. Image-based representations such as DINOv2 [35] perform poorly on this task, highlighting the limitations of appearance features for localized action recognition. Video representations such as V-JEPA2 [2] improve performance but remain constrained by their lack of explicit motion modeling. Motion-focused approaches such as DisMo [39] achieve stronger results, confirming the importance of motion representations for this task. Nonetheless, our method outperforms these baselines by a significant margin. While other methods lack access to the surrounding context due to the cropping mechanism, our model preserves the full scene context when conditioning the representation on the queried actor, allowing it to leverage contextual cues that help disambiguate motion without compromising spatial locality. We ablate the importance of context by providing our model with cropped input videos as well, effectively reducing it to the same context-reduced setting as the other methods (see **Ours**$_{\text{RGB}}$ in Tab. 4). Our method suffers from a substantial drop in performance, approaching the performance of other motion encoders in this setting, which suggests that contextual information plays a vital role for classifying localized actions. However, providing global methods with scene context by processing uncropped videos is insufficient, as motion of other actors in the scene may compromise regional understanding. To highlight the influence of distractors, we reduce our method to this setting by prompting it with a global mask rather than actor-specific ones (see **Ours**$_{\text{GLOBAL}}$ in Tab. 4). Once again, we notice a significant performance drop, indicating that either side of this trade-off is suboptimal. When not subjecting ourselves to either of these constraints, our approach performs significantly better, indicating that it does not sacrifice locality for context, but instead combines the strengths of both to enable improved action recognition.

Table 2: Human preference study. Percentage of votes for realism, motion transfer quality and locality, and the absence of visible artifacts.

Model	Backbone	#Params	Realism ↑	Transfer ↑	Fewer Artifacts ↑
ATI [1]	WAN2.1 [52]	14B	1.96	4.90	5.23
WanMove [7]	WAN2.1	14B	14.05	14.38	14.38
Tora [67]	CogVideoX [63]	5B	13.73	17.32	15.69
DisMo [39]	CogVideoX	5B	21.24	26.80	20.26
Ours	CogVideoX	5B	**49.02**	**36.60**	**44.44**

Table 3: Composable Motion Transfer. We report temporal consistency, in-region motion fidelity, out-of-region motion leakage, and motion selectivity when transferring motion from multiple source objects. Our approach substantially reduces motion leakage and achieves the best overall motion selectivity while remaining competitive in in-region motion fidelity.

Model	Temporal Consistency ↑	In-Region Motion Fidelity ↑	Out-of-Region Motion Leakage ↓	Motion Selectivity ↑
ATI [1] (Wan 2.1-14B [52])	0.9634	**0.7148**	<u>0.3857</u>	<u>0.3291</u>
WanMove [7] (Wan 2.1-14B)	<u>0.9792</u>	<u>0.7059</u>	0.4110	0.2949
Tora [67] (CogVideoX-5B [63])	0.9766	0.6520	0.3904	0.2616
Ours (CogVideoX-5B)	**0.9831**	0.6718	**0.2604**	**0.4114**

4.3 From Local to Global: Generic Motion Representations

While our framework is designed to learn spatially localized motion representations, an important question is whether such representations remain useful beyond localized queries. Increasing spatial specificity could push representations toward low-level motion signals that lose higher-level semantic properties or overly emphasize local cues. In the following, we show that neither occurs: our learned embeddings generalize naturally to standard global motion understanding tasks while retaining the semantic properties characteristic of global motion representations.

Localized Motion Pre-training Generalizes to Global Understanding. We evaluate our model in a *global zero-shot action classification* setting on a diverse set of motion-centric video benchmarks. Specifically, we evaluate on ARID [61], which measures action recognition robustness under low-light conditions; IARD [48], which evaluates identity-invariant motion understanding by minimizing reliance on appearance cues; Jester [33], a large-scale hand-gesture dataset that requires sensitivity to fine-grained local motions; Something-Something V2 [12], which focuses on subtle human-object interactions and temporal reasoning; and Diving48 [27], a challenging fine-grained action recognition benchmark designed to reduce scene and object bias, requiring recognition based primarily on body dynamics and motion patterns. Although our framework is designed to learn *localized* motion representations, these tasks require predicting a *single label for the entire scene*. To adapt our method without architectural changes, we simply provide a mask covering the full start frame when computing the content query. This effectively reduces the localized motion encoder to a *global motion encoder*, while leaving the remainder of the pipeline unchanged. The resulting performance is reported in Tab. 5. Our method achieves strong performance across all benchmarks and obtains the best results on several datasets. This demonstrates that localized motion pre-training does not limit the model to localized inference. Instead, the learned representation naturally extends to the global setting and remains highly effective for standard action recognition tasks. Importantly, our training paradigm therefore yields a *strictly more general motion encoder*. When no spatial conditioning is provided, the model behaves as a global motion encoder and achieves competitive or state-of-the-art performance. At the same time, it retains the ability to extract *region-specific motion representations* when a query mask is given. In contrast to prior approaches that are inherently restricted to global motion encoding, our model seamlessly supports both regimes within the same representation.

Spatial Localization Preserves Appearance Invariance. A key challenge for localized motion representations is that increasing spatial specificity can push representations toward low-level signals like optical flow, which capture motion locally but lose semantic invariances. Conversely, global video representations exhibit strong appearance invariance but cannot isolate individual entities' motion. Our goal is to combine both advantages: spatially localized representations that preserve the semantic invariances of global ones. Specifically, we target invariance to appearance-driven transformations, namely viewpoint changes, dynamic attribute modifications, entity replacement across semantic classes, and stylistic rendering changes, since these are critical for transferability to novel structures.

Table 4: Localized Motion Classification. Left: Comparison of image, video, and motion encoders for localized action classification. Our method outperforms global approaches that obtain localization by cropping the video input (RGB) or masking features in latent space (GLOBAL). Ablations show that removing context through RGB crops (**Ours**$_{\text{RGB}}$) or removing localization by querying the entire frame (**Ours**$_{\text{GLOBAL}}$) both degrade performance. The best results are achieved when motion is encoded locally while preserving scene context. **Right:** Example of a positive and negative classification outcome with regard to a reference subject from the A2D dataset.

	Accuracy↑	F1↑
Image encoders		
DINOv2$_{\text{RGB}}$ [35]	10.9	17.4
DINOv2$_{\text{LATENT}}$	6.0	10.2
Video encoders		
V-JEPA2$_{\text{RGB}}$ [2]	14.6	23.8
V-JEPA2$_{\text{LATENT}}$	13.9	21.6
Global motion encoders		
DINOv2$_{\text{RGB}}$ + SemanticMoments [19]	12.4	18.8
DINOv2$_{\text{LATENT}}$ + SemanticMoments	5.5	9.5
V-JEPA2$_{\text{RGB}}$ + SemanticMoments	14.3	23.4
V-JEPA2$_{\text{LATENT}}$ + SemanticMoments	11.9	18.8
DisMo$_{\text{RGB}}$ [39]	22.4	33.9
Ours$_{\text{RGB}}$	20.0	31.1
Ours$_{\text{GLOBAL}}$	23.4	35.1
Ours	**29.9**	**42.0**

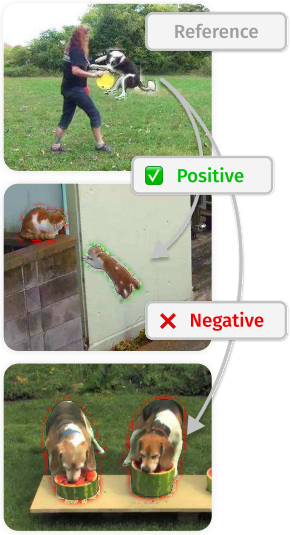


Table 5: Zero-shot action classification in a global setting. We report top-1 accuracy on several motion-centric video benchmarks using a *k*NN classifier with *k=20*. Although our model is trained to produce localized motion representations, providing a full-frame query reduces it to a global motion encoder, yielding strong performance across benchmarks.

	SSv2↑	Jester↑	Diving48↑	ARID↑	IARD↑
DINOv2 [35]	10.3	23.2	9.39	14.5	76.1
VideoMAE [50]	7.1	20.1	-	17.3	73.4
V-JEPA2 [2]	22.2	40.8	11.1	28.0	87.6
SemanticMoments$_{\text{DINO}}$ [19]	11.4	37.3	9.85	22.0	62.7
SemanticMoments$_{\text{V-JEPA2}}$ [19]	**31.5**	52.2	12.7	38.3	<u>92.5</u>
DisMo [39]	24.6	<u>69.8</u>	**20.9**	<u>55.3</u>	92.0
Ours	<u>26.5</u>	**72.8**	<u>17.1</u>	**55.4**	**93.0**

Table 6: Motion retrieval under appearance transformations. We report top-1 retrieval accuracy on the SemanticMoments synthetic split averaged over all augmentations.

	Accuracy↑
DINOv2 [35]	44.4
VideoMAE [50]	79.2
V-JEPA2 [2]	74.4
SemanticMoments$_{\text{DINO}}$ [19]	86.4
SemanticMoments$_{\text{V-JEPA2}}$ [19]	84.4
DisMo [39]	**90.0**
Ours	<u>89.6</u>

We evaluate on the synthetic split of the *SemanticMoments* benchmark[19], which measures motion retrieval under controlled appearance transformations (added objects, appearance changes, object replacement, viewpoint variation, style changes) using a top-1 retrieval protocol. We compare against DINOv2[35], V-JEPA2[2], and DisMo[39], applying the benchmark's temporal aggregation to DINOv2 and V-JEPA2. Results are shown in Tab. 6. These show our method outperforms all baselines except DisMo, achieving near-parity while substantially surpassing appearance-centric representations. This demonstrates that spatial localization does not push representations toward low-level motion cues; instead, it preserves semantically meaningful motion dynamics while enabling localized reasoning.

Action Probing under Distractors. Following MaskLAM [10], we test action alignment under visual distractors using the public LAOM DCS dataset [46, 51]

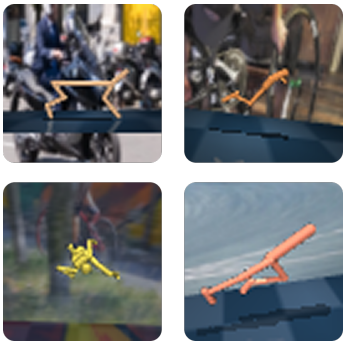


Figure 6: DCS examples. The dataset introduces visual distractors through randomized agent colors, video backgrounds, and camera motion.

(for examples see Fig. 6). We freeze each representation and train only a linear action probe to predict the actions. Our localized LAM obtains the lowest mean NMSE, improving over uncropped DisMo by 32.3%, while direct RGB masking also substantially improves DisMo but remains weaker than our representation. These results show that spatial localization improves the robustness and action alignment of motion representations under visual distractors.

Table 7: Frozen action probing under visual distractors. We report normalized MSE (NMSE; lower is better) on the public LAOM DCS dataset using a linear action probe on frozen representations. Localizing the motion representation substantially improves over global DisMo and also outperforms RGB cropping, indicating that representation-level localization better preserves action-relevant information under distractors. MaskLAM results are taken from [10] and use a larger training dataset.

	NMSE ↓				
	Cheetah Run	**Hopper Hop**	**Humanoid Walk**	**Walker Run**	**Mean**
MaskLAM [10]	0.1148	0.2601	0.3852	0.1704	0.2326
DisMo [39]	0.6095	0.7250	0.6870	0.5446	0.6415
DisMo (RGB cropped) [39]	0.4201	0.5406	0.5225	0.5138	0.4993
Ours	**0.4161**	**0.4744**	**0.4605**	**0.3845**	**0.4339**

5 Conclusion

In this work, we introduce a promptable and localized motion representation that encodes region-specific dynamics while preserving full-scene context. We showed that local motion cannot be reliably inferred from spatial crops, which discard context, nor from post-hoc feature masking, which does not condition the encoding process on the queried region. By conditioning motion encoding directly on spatial masks while processing the full video, our approach produces temporally consistent and spatially precise motion embeddings. Across motion transfer and localized classification tasks, we demonstrated improved controllability, structured scene composition, and stronger motion understanding compared to global video representation localized via cropping or feature masking. We achieve these results, while still demonstrating strong performance on global video understanding tasks, such as classification and retrieval. Our results highlight the importance of encoding local motion while preserving full-scene context for structured and compositional video modeling.

Acknowledgment

This project has been supported by the Horizon Europe project ELLIOT (GA No. 101214398), the project "GeniusRobot" (01IS24083) funded by the Federal Ministry of Research, Technology and Space (BMFTR), the BMWE ZIM-project (No. KK5785001LO4) "conIDitional LoRA", the German Federal Ministry for Economic Affairs and Energy within the project "NXT GEN AI METHODS - Generative Methoden für Perzeption, Prädiktion und Planung", the bidt project KLIMA-MEMES, and the German Academic Exchange Service (DAAD) under the Kondrad Zuse School of Excellence for Reliable AI (RelAI). The authors gratefully acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS/JUPITER at JSC and the HPC resources supplied by the NHR@FAU Erlangen. We thank Timy Phan and for their helpful feedback and support, and Owen Vincent for continuous technical support.

References

[1] Wang Angtian, Huang Haibin, Fang Zhiyuan, Yang Yiding, and Ma Chongyang. ATI: Any trajectory instruction for controllable video generation. *arXiv preprint*, arXiv:2505.22944, 2025.

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[3] 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.

[4] 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.

[5] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise, 2025.

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

[7] Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, Yingya Zhang, and Yujiu Yang. Wan-move: Motion-controllable video generation via latent trajectory guidance, 2025.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023.

[10] Marcus Fechner, Hamza Adnan, Constantin C Lüth, Matthew T Jackson, Alexey Zakharov, and J Marius Zöllner. Segment to focus: Guiding latent action models in the presence of distractors. *arXiv preprint arXiv:2602.02259*, 2026.

[11] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024.

[12] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *CVPR*, pages 5842–5850, 2017.

[13] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025.

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2024.

[15] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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

[17] 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. arxiv 2021. *arXiv preprint arXiv:2106.09685*, 2021.

[18] 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, 2021.

[19] Saar Huberman, Kfir Goldberg, Or Patashnik, Sagie Benaim, and Ron Mokady. Semanticmoments: Training-free motion similarity via third moment features. *arXiv preprint arXiv:2602.09146*, 2026.

[20] Saar Huberman, Kfir Goldberg, Or Patashnik, Sagie Benaim, and Ron Mokady. Semanticmoments: Training-free motion similarity via third moment features, 2026.

[21] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8079–8088, 2024.

[22] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *CVPR*, pages 9212–9221, 2024.

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

[24] 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.

[25] Mathis Koroglu, Hugo Caselles-Dupré, Guillaume Jeanneret, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6226–6236, 2025.

[26] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[27] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European conference on computer vision (ECCV)*, pages 513–528, 2018.

[28] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024.

[29] Penghui Liu, Jiangshan Wang, Yutong Shen, Shanhui Mo, Chenyang Qi, and Yue Ma. Multimotion: Multi subject video motion transfer via video diffusion transformer, 2025.

[30] Yanchen Liu, Yanan Sun, Zhening Xing, Junyao Gao, Kai Chen, and Wenjie Pei. Motionshot: Adaptive motion transfer across arbitrary objects for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11861–11871, 2025.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[32] Yue Ma, Zhikai Wang, Tianhao Ren, Mingzhe Zheng, Hongyu Liu, Jiayi Guo, Mark Fong, Yuxuan Xue, Zixiang Zhao, Konrad Schindler, et al. Fastvmt: Eliminating redundancy in video motion transfer. *arXiv preprint arXiv:2602.05551*, 2026.

[33] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[34] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *ICLR*, 2025.

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2023. arXiv:2304.07193 [cs].

[36] Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers, 2025.

[37] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation, 2018.

[38] 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, 2024.

[39] Thomas Ressler-Antal, Frank Fundel, Malek Ben Alaya, Stefan Andreas Baumann, Felix Krause, Ming Gui, and Björn Ommer. Dismo: Disentangled motion representations for open-world motion transfer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[40] 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 (CVPR)*, pages 10684–10695, 2022.

[41] Mingzhi Sheng, Zekai Gu, Peng Li, Cheng Lin, Hao-Xiang Guo, Ying-Cong Chen, and Yuan Liu. Flexam: Flexible appearance-motion decomposition for versatile video generation control. *arXiv preprint arXiv:2602.13185*, 2026.

[42] Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Xiangtai Li, and Yunhai Tong. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. *arXiv preprint arXiv:2503.17350*, 2025.

[43] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[44] Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Schechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls, 2025.

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[46] Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite - a challenging benchmark for reinforcement learning from pixels. *ArXiv*, abs/2101.02722, 2021.

[47] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

[48] Andrea Tacchetti, Leyla Isik, and Tomaso Poggio. Invariant Action Recognition Dataset, 2019.

[49] Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer, 2025.

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

[51] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020.

[52] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025.

[53] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024.

[54] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[55] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024.

[56] Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. *arXiv preprint arXiv:2411.19324*, 2024.

[57] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[58] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? action understanding with multiple classes of actors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2264–2273, 2015.

[59] Chenliang Xu, Caiming Xiong, and Jason J. Corso. Action understanding with multiple classes of actors, 2017.

[60] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark, 2018.

[61] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2*, pages 70–84. Springer, 2021.

[62] 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.

[63] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025.

[64] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *CVPR*, pages 8466–8476, 2024.

[65] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[66] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023.

[67] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025.

[68] Zhenghao Zhang, Junchao Liao, Xiangyu Meng, Long Qin, and Weizhi Wang. Tora2: Motion and appearance customized diffusion transformer for multi-entity video generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9434–9443, 2025.

[69] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024.

[70] Artem Zholus, Carl Doersch, Yi Yang, Skanda Koppula, Viorica Patraucean, Xu Owen He, Ignacio Rocco, Mehdi S. M. Sajjadi, Sarath Chandar, and Ross Goroshin. Tapnext: Tracking any point (tap) as next token prediction, 2025.

[71] Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. Trackgo: A flexible and efficient method for controllable video generation, 2025.

Table A.1: Pretraining hyperparameters for our model. Summary of training data, preprocessing, optimization settings, and architecture configuration.

Category	Component	Setting
Data	Frames per video	8
	Frame resolution	256×256
	Tracks per decoding step	256
	Masks per video	4
Transformations	Random resized crop	(0.04, 1.0)
	Rotation	$(-30, 30)$
	Aspect ratio	$(3/4, 4/3)$
	Brightness	(0.5, 1.5)
	Contrast	(0.5, 1.5)
	Saturation	(0.5, 1.5)
Optimization	Batch size	64
	Training schedule	600k iterations
	Optimizer	AdamW
	Warmup iterations	5000
	lr scheduler	constant w/ warmup
	lr	5×5^{-5}
	Weight decay	0.01
Architecture	Frame embedder	ViT-Base (12 layers, dim 768)
	Frame embedder init	DINOv2 pretrained weights
	Content encoder	4-layer cross-attention aggregator
	Content query tokens	1 per entity
	Content embedding dim	768
	Motion encoder	12-layer Vision Transformer
	Motion queries	Content embeddings repeated across frames
	Tracks decoder	12-layer Vision Transformer
	Tracks decoder conditioning	Adaptive RMSNorm (motion + content)

Supplementary Material

A Implementation Details

A.1 Pretraining details

Our architecture consists of four main components: a *frame embedder*, a *content encoder*, a *motion encoder*, and a *decoder*. All transformer-based modules operate with a hidden dimension of 768. Details about training data, optimization, and method architecture can be found in Supp. Tab. A.1.

Frame Embedder. Input video frames are first processed by a shared frame embedding network that produces per-frame token embeddings. The frame embedder is implemented as a 12-layer Vision Transformer with hidden dimension 768. To accelerate convergence and leverage strong visual representations, the model is initialized with weights from [35]. The resulting frame tokens serve as input to all subsequent modules.

Content Encoder. The content encoder provides an entity-specific appearance representation that identifies the region whose motion should be modeled. Given a video frame and a corresponding entity mask, it acts as a mask-conditioned region encoder that maps the visual content of the selected entity to a single content embedding. This formulation is independent of the particular aggregation architecture: the region encoder can, for example, be implemented through masked pooling, mask-conditioned attention over image tokens, or an image encoder that receives the mask as an additional conditioning signal. For the experiments in this work, we instantiate the content encoder as a masked attentive pooling network over spatial frame embeddings. A learnable query token attends to the frame embeddings through four cross-attention layers, where the entity mask is used to bias the aggregation toward the target region. Each cross-attention layer is followed by an MLP block for further feature aggregation. The resulting query token forms the *content embedding* of the entity.

Motion Encoder. The motion encoder produces motion embeddings for each entity across the temporal dimension of

the video. It takes as input the frame embeddings of all frames together with the entity-specific content embeddings. For each entity, the corresponding content embedding is repeated across the temporal dimension and used as a query token. These queries interact with the frame embeddings through a 12-layer Vision Transformer with hidden dimension 768, resulting in *motion embeddings* that describe the entity-specific motion at each frame.

Decoder. The decoder predicts motion in the form of point trajectories relative to a source frame. For each entity, we sample a set of starting points within the corresponding entity region of the source frame and obtain their ground-truth trajectories across the video using TAPNext [70]. The decoder is then trained to predict these temporal trajectories for a fixed number of prediction steps. The decoder receives the frame embeddings of the source frame together with the corresponding motion and content embeddings. It is implemented as a 12-layer Vision Transformer with hidden dimension 768. Conditioning on motion and content embeddings is implemented using *adaptive RMSNorm*, allowing the decoder activations to be modulated by the motion and content signals. The decoder outputs the predicted trajectories for the sampled points, expressed as displacements relative to their initial positions in the source frame.

Training Data. We train on a mixture of two datasets covering approximately 17M video clips. The first one is OpenVid-1M [34], a high-quality open-domain dataset well suited for generative tasks. The second one is a self-collected dataset of approximately 16M clips sourced from YouTube, covering a broad range of real-world scenes and motions. We provide an overview of the dataset composition in Supp. Tab. A.2.

Table A.2: Training data composition.

Dataset	Domain	# Clips
OpenVid-1M [34]	open	1M
Internal	open	16M
Combined	open	17M

A.2 Video Model Finetuning Details

Training Setup. We adapt the attention and feedforward layers of CogVideoX-5B [63] using LoRA [17] modules with rank 128. We fine-tune the model using a batch size of 16 and the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and learning rate 5×10^{-5} with a constant schedule and 250 linear warmup steps. Training is performed in `bfloat16` mixed precision. Videos are processed at a resolution of $25 \times 480 \times 720$ (frames $\times$ height $\times$ width) and generated at 8 FPS.

Temporal Alignment of Motion Conditioning. The CogVideoX VAE compresses T=25 input frames into N_ℓ=7 latent frames using a temporal compression ratio of c_t=4. Our motion representation model extracts embeddings via a sliding window of size p+1 with prediction horizon p=3, producing $T - (p{+}1) = 21$ motion embeddings per region. To align these with the VAE latent frames, we prepend $p = 3$ learned padding tokens, yielding $T - 1 = 24$ temporally contiguous embeddings. These are grouped into non-overlapping chunks of c_t=4, producing $(T{-}1)/c_t = 6$ temporally aligned representations. Each chunk is concatenated along the feature dimension and projected through a learned mapping network to produce a single conditioning vector per latent frame. A final learned padding vector is appended for the last latent frame, yielding N_ℓ=7 conditioning vectors in one-to-one correspondence with the VAE latent frames. Motion and content tokens are concatenated along the channel dimension before this projection, allowing the model to condition on joint motion–content representations while keeping the base weights frozen.

Multi-region Conditioning. For multi-region scenarios, the above alignment procedure is applied independently per region mask, producing $K \times N_\ell$ conditioning tokens for K regions. All per-region tokens are concatenated along the sequence dimension and appended to the self-attention layers of the transformer. To preserve spatial and temporal grounding, we apply 3D rotary positional embeddings (RoPE) [47] to all conditioning tokens. The temporal position of each token is determined by its corresponding latent frame index, while the spatial position is set to the centroid of its associated region mask. This ensures that each motion token attends with positional bias consistent with the spatiotemporal location of its corresponding region in the generated video.

A.3 Evaluation Details

Datasets. We evaluate on a diverse set of benchmarks spanning action classification and retrieval, each emphasizing different aspects of temporal and spatial reasoning. Something-Something v2 (SSv2) [12] is a large-scale dataset of human–object interactions designed to test fine-grained temporal reasoning while minimizing reliance on background context or object identity. Jester [33] contains over 148,000 short clips of hand gestures, serving as a benchmark for fine-grained, spatially localized motion sensitivity. ARID [61] (Action Recognition in the Dark) comprises 5,500 RGB-D videos captured under low-light conditions, evaluating robustness to appearance degradation. IARD [48] (Invariant Action Recognition Dataset) consists of controlled recordings of five actors performing five actions from multiple viewpoints, enabling evaluation of identity- and viewpoint-invariant motion recognition. Diving48 [26] contains approximately 18,000 trimmed clips of competitive diving covering 48 standardized dive sequences, requiring modeling of long-term temporal dynamics across takeoff, flight, and entry stages. A2D [59] provides 3,782 videos with pixel-level actor and action annotations, enabling evaluation of spatially grounded action understanding across multiple actor–action combinations. The SemanticMoments synthetic benchmark [19] provides a controlled evaluation setting designed to isolate specific failure modes of video representations. By holding the underlying motion constant while systematically varying non-motion factors such as appearance, object identity, and viewpoint, it enables precise analysis of how these confounds influence representation similarity.

Motion Transfer using Track-conditioned Models. To perform motion transfer using track-conditioned video generation baselines, we re-target point trajectories from a source video onto a target image before conditioning the generator. Given a source video with associated object masks, we first extract dense point tracks within the masked region using CoTracker3 [23]. These tracks are then spatially transferred to align with the target objects by computing an affine mapping between the bounding boxes of the source and target masks: each track point is translated so that the source mask centroid coincides with the target mask centroid, and uniformly scaled so that the source bounding box fits within the target bounding box. The retargeted tracks, together with the target image, are then provided as conditioning inputs to the track-conditioned video generation model.

B Ablations

In this section we present ablation studies that analyze several architectural and objective design choices of our model. These experiments aim to justify the configuration used in the final model and provide insight into how different components affect the learned motion representations.

Specifically, we investigate three aspects: (i) the prediction distance used during trajectory decoding, (ii) the dimensionality of the motion embedding bottleneck, and (iii) the contribution of the content encoder and its conditioning during trajectory decoding.

B.1 Prediction Distance

We study the effect of the *prediction distance* p used during trajectory decoding, which determines how many steps into the future point trajectories are being predicted relative to their starting positions in the source frame. At one extreme, $p = 1$ reduces to sparse optical flow estimation; at the other, $p = 7$ requires predicting trajectories across the entire video (8 frames total). We train seven model variants with $p \in \{1, \ldots, 7\}$, each for 300k steps, and evaluate on six action classification benchmarks: A2D, Something-Something-V2 (SSv2), Jester, Diving48, IARD, and ARID. A2D is a localized multi-label benchmark whereas the remaining ones are global datasets with a single label per video clip. Results are summarized in Supp. Fig. B.1.

Overall, the results indicate that our method is robust to the choice of prediction distance. While SSv2 and Jester exhibit a moderate performance drop beyond $p = 3$, the differences remain small, suggesting that the prediction distance is not a critical design choice. We adopt $p = 3$ as a practical default based on these results, though nearby values yield comparable performance.

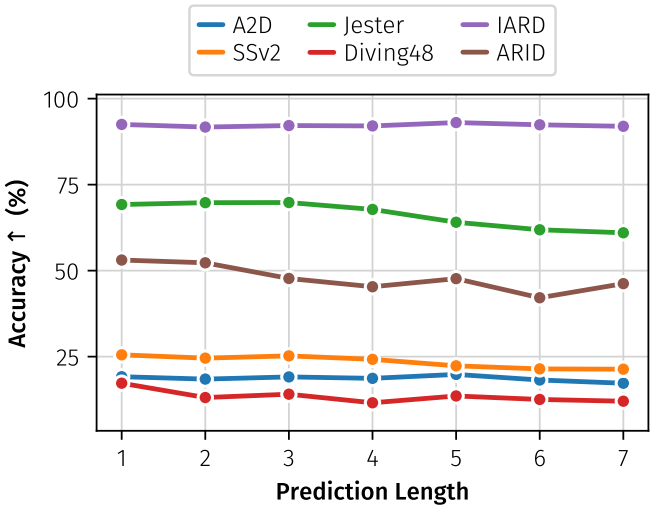


Figure B.1: Action classification accuracy across varying prediction lengths p. Performance remains stable across $p \in \{1, \ldots, 7\}$, with minor degradation beyond $p = 3$, suggesting that our framework is robust to different choices of prediction length.

B.2 Motion Bottleneck Size

We study the effect of the motion embedding dimensionality on downstream performance. We train model variants with bottleneck sizes ranging from $d = 128$ to $d = 768$ and evaluate on the same action classification benchmarks used in Sec. B.1. We adopt $d = 768$ as the default in all main experiments. Results are summarized in Supp. Fig. B.2.

Across all benchmarks, performance remains remarkably stable even as the bottleneck size is reduced by a factor of six, from $d = 768$ down to $d = 128$. This indicates that the motion dynamics captured by our representation can be effectively compressed into compact embeddings without significant information loss. The robustness to bottleneck size suggests that even more aggressive compression may be feasible, making the approach well suited for downstream applications where compact representations are desirable.

B.3 Appearance Entanglement Analysis

We analyze the degree to which the learned motion embeddings remain entangled with appearance information. Ideally, motion representations should encode *how* an entity moves while remaining independent of *who* performs the motion.

To quantify appearance leakage, we evaluate motion embeddings on the IARD dataset by measuring actor identity classification with a kNN classifier. The dataset contains five actors; therefore, a representation that is fully independent of appearance should achieve approximately 20% accuracy (chance level). Higher accuracy indicates that identity information remains encoded in the motion embedding. We compare two model variants that differ in how the decoder is conditioned during track reconstruction. Importantly, both variants already receive implicit appearance context: tracks are only queried at spatial locations that lie *inside the masked regions*. Consequently, the decoder is naturally conditioned on the localized region of the entity, which already encourages a degree of appearance disentanglement. The difference between the two variants lies in whether the decoder additionally receives the corresponding *content token*. In the *No Content Attachment* variant, reconstruction relies solely on the motion embedding and the spatial mask constraint. In the *Content Attachment* variant, the decoder is additionally conditioned on the content token, providing a dedicated pathway for appearance information.

Figure B.3 shows the resulting actor identity classification accuracy of the motion embeddings. Even without explicit content attachment, identity predictability is already reduced due to the spatial mask conditioning. However, providing the content token during decoding further decreases identity classification accuracy, pushing performance closer to the independence line. This demonstrates that explicit content attachment strengthens the separation between motion and

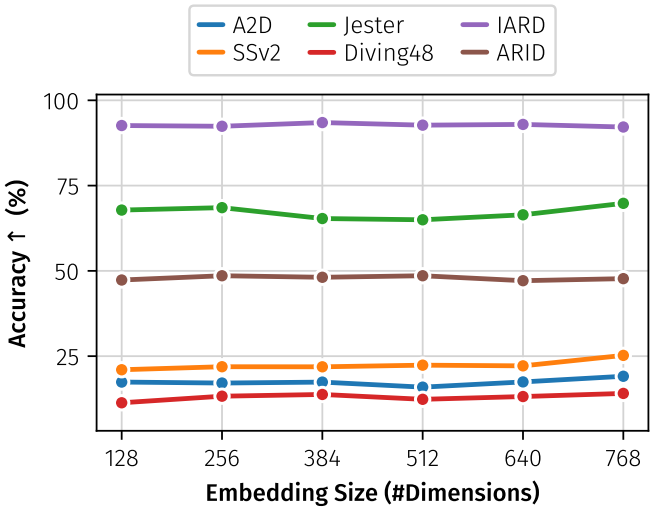


Figure B.2: Action classification accuracy across motion bottleneck sizes. Performance remains stable from $d = 768$ down to $d = 128$, indicating that motion dynamics are effectively captured even with compact embeddings.

appearance, encouraging motion embeddings that are more appearance-invariant.

To further assess appearance leakage, we train a flow-matching decoder to reconstruct the input frame from frozen motion embeddings. Since the encoder cannot adapt during probe training, better reconstruction indicates that more appearance information is retained in its representation. As shown in Supp. Tab. B.3 and Supp. Fig. B.4, our embeddings yield the highest L1 and LPIPS errors and the lowest CLIP similarity, indicating that they contain less recoverable appearance information than the baseline representations.

C Qualitative Examples

We present additional qualitative results for motion transfer and compare them against the previously introduced baselines. The results cover both one-to-one transfer and compositional settings with multiple motion sources mapped to multiple target objects. All experiments are conducted using our adapted CogVideoX-5B model.

C.1 Localized Motion Transfer

Qualitative examples for the one-to-one motion transfer setting are shown in Supp. Figs. C.5 to C.9. Prior methods exhibit several consistent failure modes in localized motion transfer. Low-level motion transfer approaches struggle under pose mismatch and often produce geometric inconsistencies and strong visual artifacts, such as clipping or abrupt scene changes. They also tend to entangle local object motion with global scene motion, causing subject motion to be mistaken for camera motion and vice versa. Global motion transfer methods instead fail to properly localize motion, frequently moving the wrong subject, affecting multiple instances, or hallucinating additional objects. Our method is substantially more robust to these issues, producing cleaner transfers that remain localized to the correct subject.

C.2 Scene Composition

Qualitative examples for the compositional motion transfer setting, where multiple motion sources are mapped to multiple target objects, are shown in Supp. Figs. C.10 and C.12 to C.14. We utilize low-level methods as a comparison for the composite motion transfer examples, global motion transfer models are not applicable, since their video generation pipeline does not accept multiple motion embeddings from different subjects. In composite motion transfer, prior methods mainly fail by misassigning motions across subjects, omitting parts of the requested motion composition, or introducing degenerate camera motion. These errors indicate difficulty in jointly modeling multiple subject-specific motions while preserving a stable global scene. As a result, the generated videos often do not reflect the intended

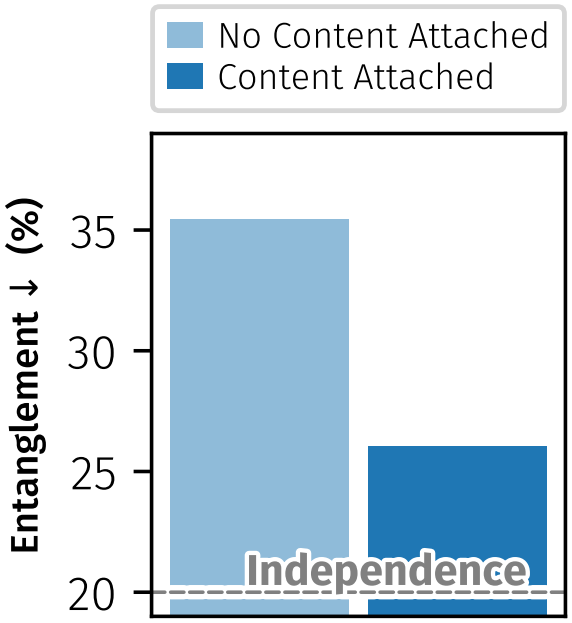


Figure B.3: Appearance entanglement analysis. We quantify appearance leakage in motion embeddings on IARD by measuring actor identity classification using a kNN classifier. With five actors, perfect independence corresponds to 20% accuracy. **Left:** Model without decoder conditioning on the content token. **Right:** Model with content-token conditioning. Conditioning markedly reduces identity predictability toward the independence line, highlighting the role of content attachment in disentangling motion from appearance.

Table B.3: Appearance reconstruction probe. Lower CLIP similarity and higher L1 and LPIPS errors indicate less recoverable appearance information.

Model	CLIP ↓	L1 ↑	LPIPS ↑
V-JEPA2 [28]	0.56	0.19	0.61
DisMo [32]	**0.53**	0.26	0.67
Ours	**0.53**	**0.27**	**0.69**

composition of motions. Our method better separates subject-level motion from global scene dynamics and therefore transfers the correct motions to the correct subjects more reliably.

C.3 Text-prompted Generation

While our primary experiments focus on motion transfer conditioned on a target start frame, the learned motion representation is not restricted to image-conditioned generation. We further demonstrate that it can support promptable text-only generation by fine-tuning the video model without start-frame conditioning. In this setting, the model receives a motion embedding extracted from a driving video together with a textual description of the desired scene. The text prompt determines the generated appearance and semantic content, while the motion embedding specifies the temporal dynamics of the generated video. As shown in Supp. Fig. C.15, the model transfers the articulated motion of the driving sequence to a novel subject and scene specified only through text. This demonstrates that the representation captures reusable motion information that can be combined with different appearance-conditioning modalities, extending its applicability beyond image-to-video motion transfer.

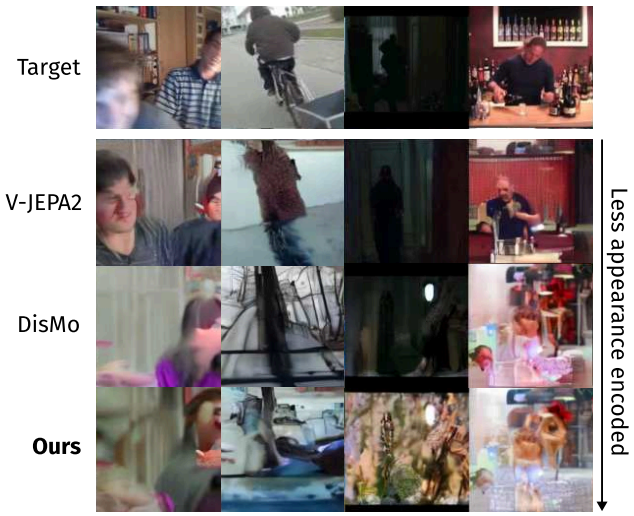


Figure B.4: Appearance reconstruction probe examples. Reconstructions from our motion embeddings preserve less of the actor's appearance than those obtained from the baseline representations.

C.4 User Study Details

We conducted a blinded user study on Prolific.com involving 18 participants and 17 non-cherrypicked test examples. For each example, participants were shown the generated videos without information about the underlying method or model identity. They independently evaluated each video along three dimensions: overall visual realism, the quality and spatial locality of the transferred motion, and the presence of visible artifacts. The transfer quality and locality criterion assessed both how faithfully the source motion was reproduced and whether it remained confined to the intended target subject without affecting unrelated regions of the scene.

C.5 Ethical Impact

As with most generative models, our approach is inherently dual-use. While it enables fine-grained control over object motion for creative, scientific, and educational applications, the same capabilities could be misused to fabricate deceptive or misleading video content. In particular, localized motion transfer may facilitate the manipulation of how individual subjects appear to move or act within a scene, potentially creating convincing visual evidence of events that never occurred. Responsible deployment should therefore include clear disclosure of generated content, provenance tracking where possible, and appropriate safeguards against deceptive use.

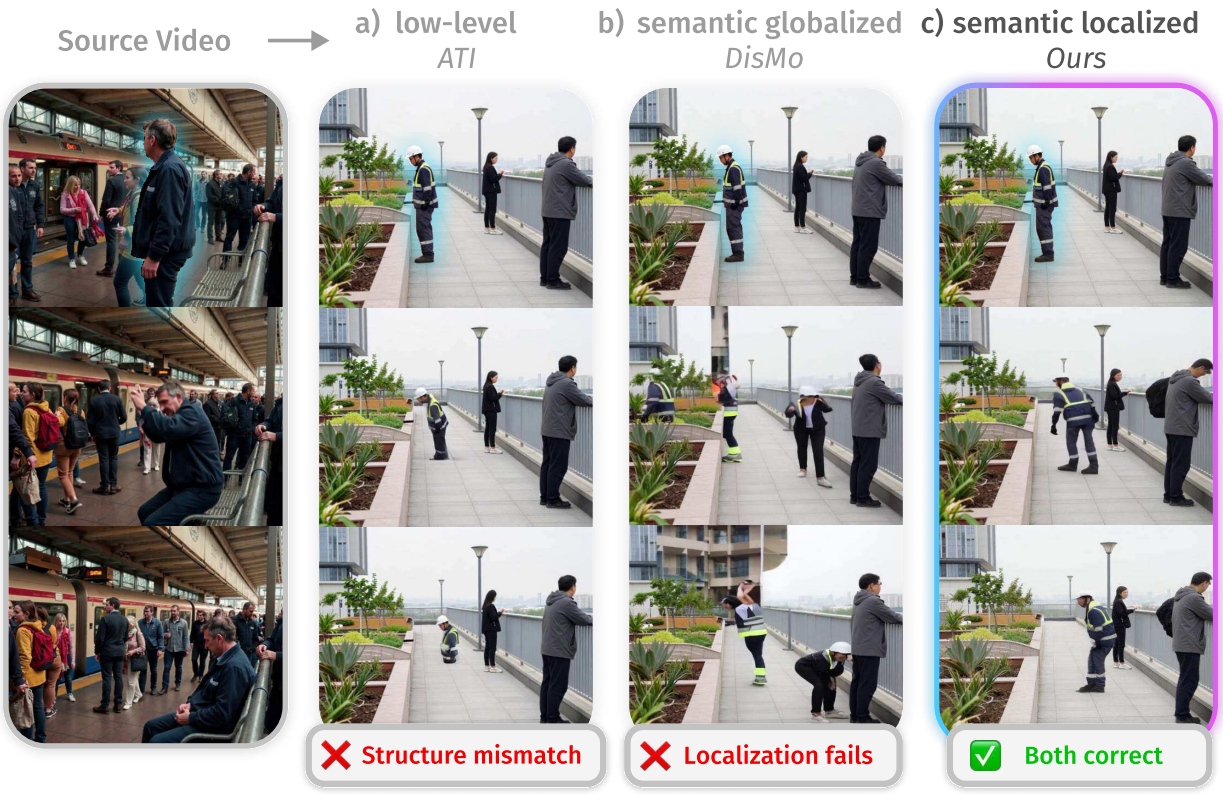


Figure C.5: Localized Motion Transfer. Poses don't align, so instead of making the person sit, **(a)** makes the person clip through the ground. **(b)** moves the wrong person, **(c)** makes the correct person sit.

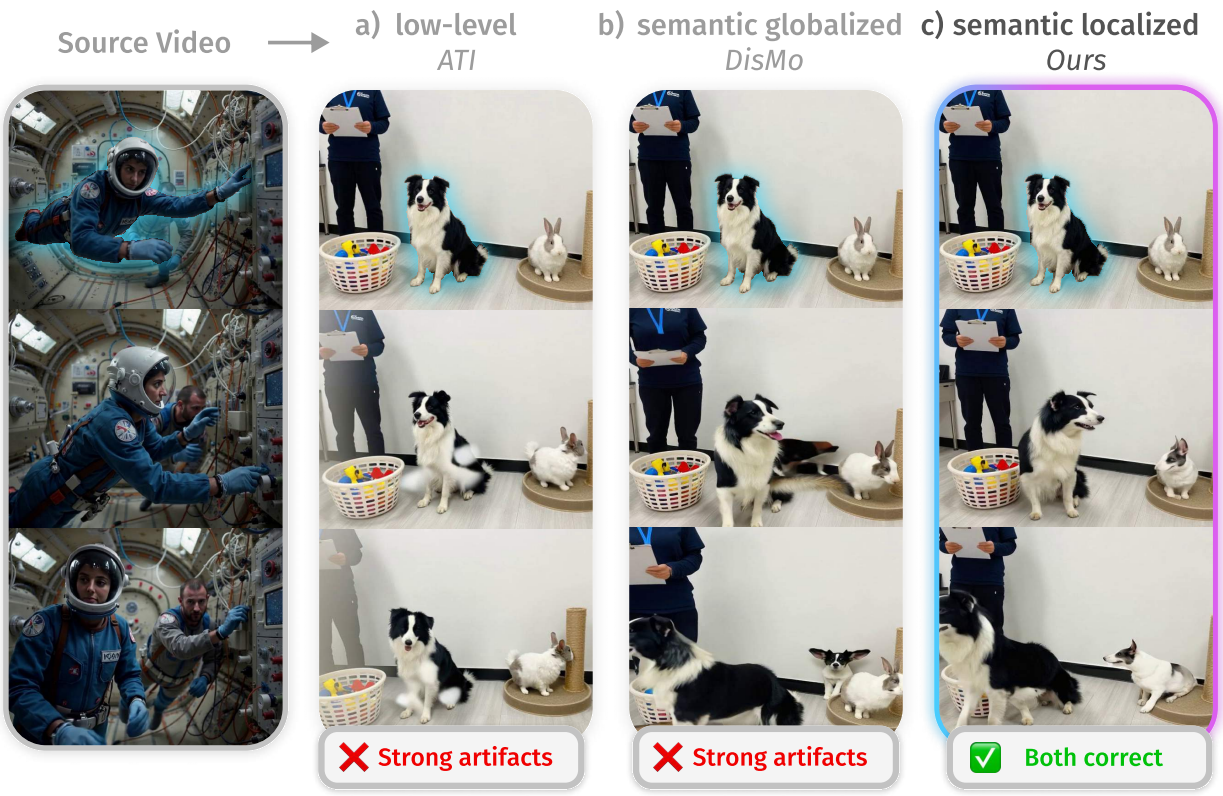


Figure C.6: Localized Motion Transfer. Since the poses don't align, **(a)** generates strong artifacts, while **(b)** generates a second dog to accommodate for the two subjects in the source video, while **(c)** only moves the correct subject without strong artifacts.

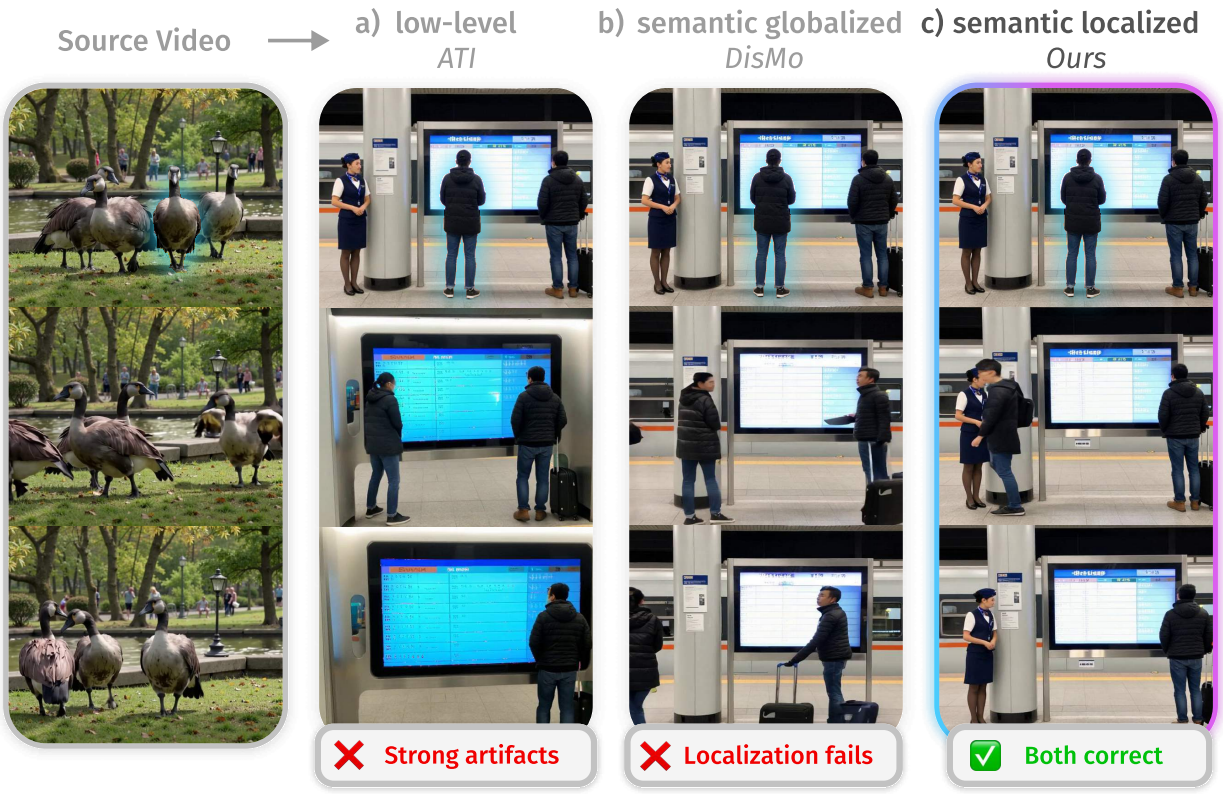


Figure C.7: Localized Motion Transfer. Since the poses don't align, **(a)** generates strong artifacts, while **(b)** moves all subjects, however **(c)** moves only the correct person.

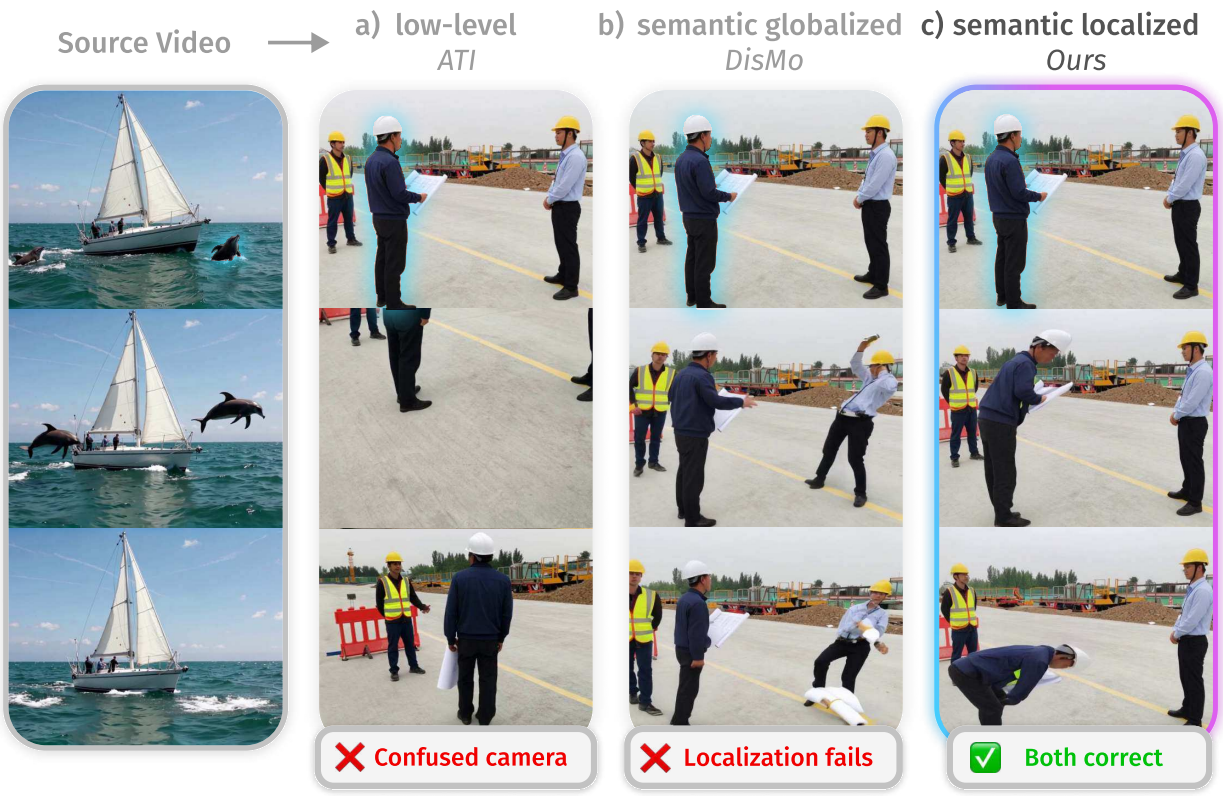


Figure C.8: Localized Motion Transfer. (a) Confuses the object motion for camera motion and does not generate the correct motion since the poses don't align, **(b)** moves the wrong person with the wrong motion, **(c)** moves the correct person with the right motion.

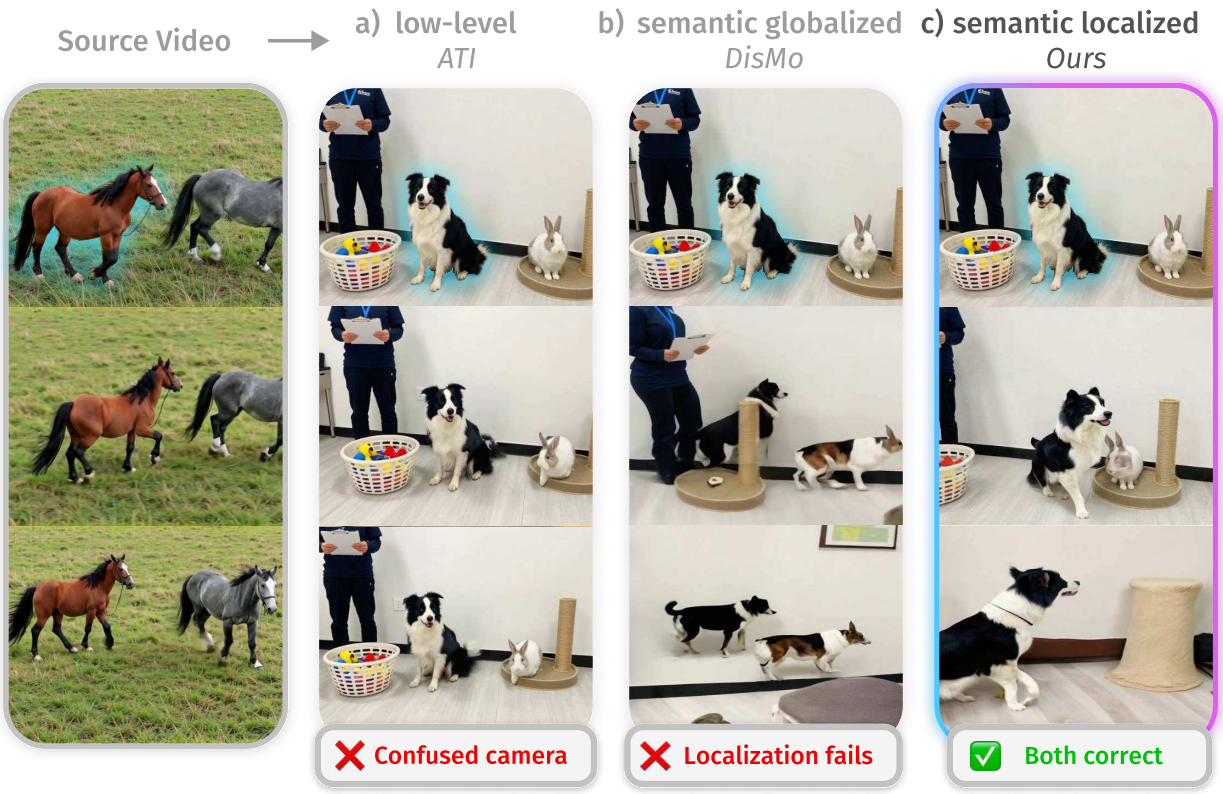


Figure C.9: Localized Motion Transfer. Since the camera follows the subject, **(a)** confuses this for static subjects, **(b)** moves both subjects unrealistically, while **(c)** moves the correct subject with the right motion.



Figure C.10: Composite Motion Transfer. (a) creates degenerate camera motion and does not transfer the correct motions **(b)** transfers the correct motion to the right subjects.

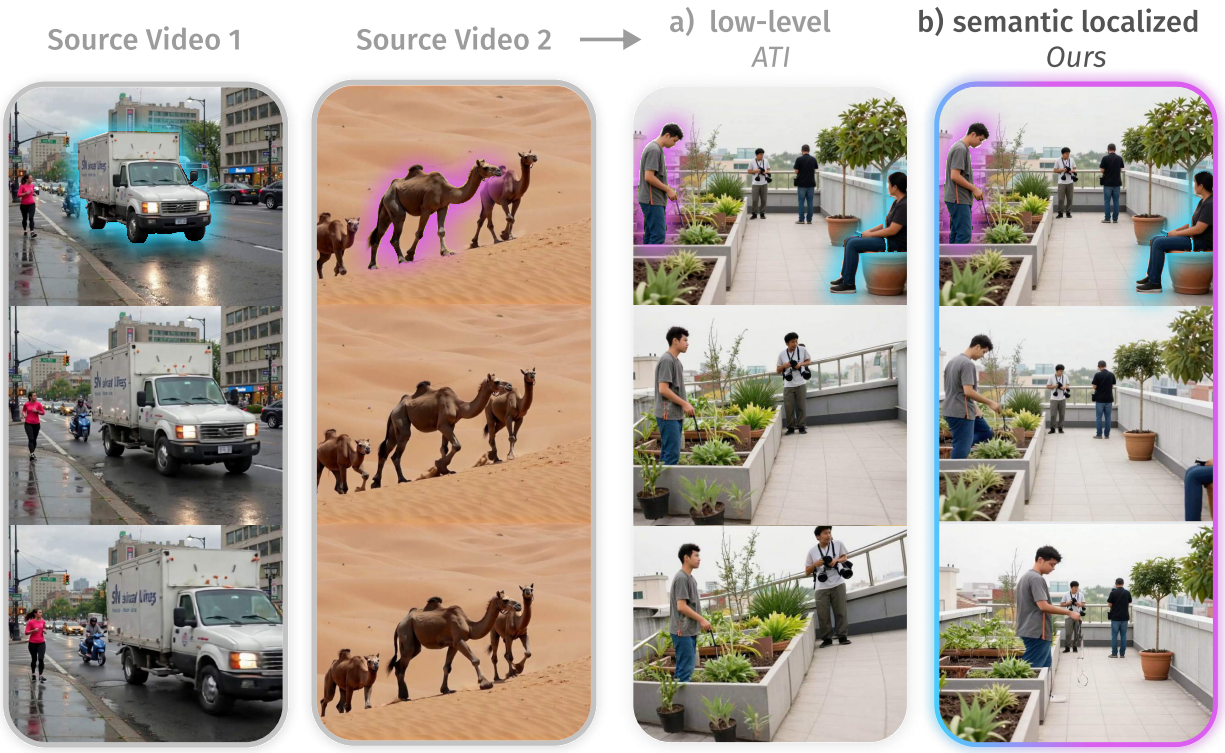


Figure C.11: Composite Motion Transfer. (a) creates degenerate camera motion and does not transfer the correct motions, while **(b)** transfers the correct motion to the right subjects.

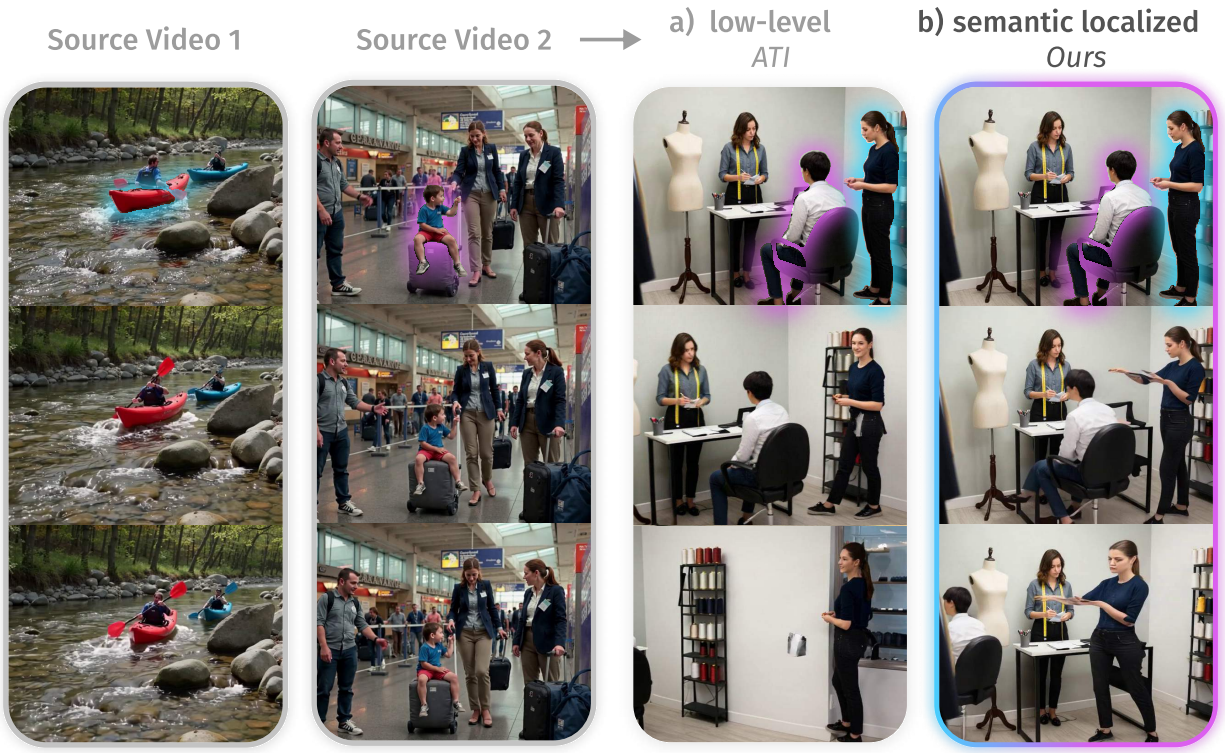


Figure C.12: Composite Motion Transfer. (a) creates degenerate camera motion and does not transfer the correct motions, while **(b)** transfers the correct motion to the right subjects.

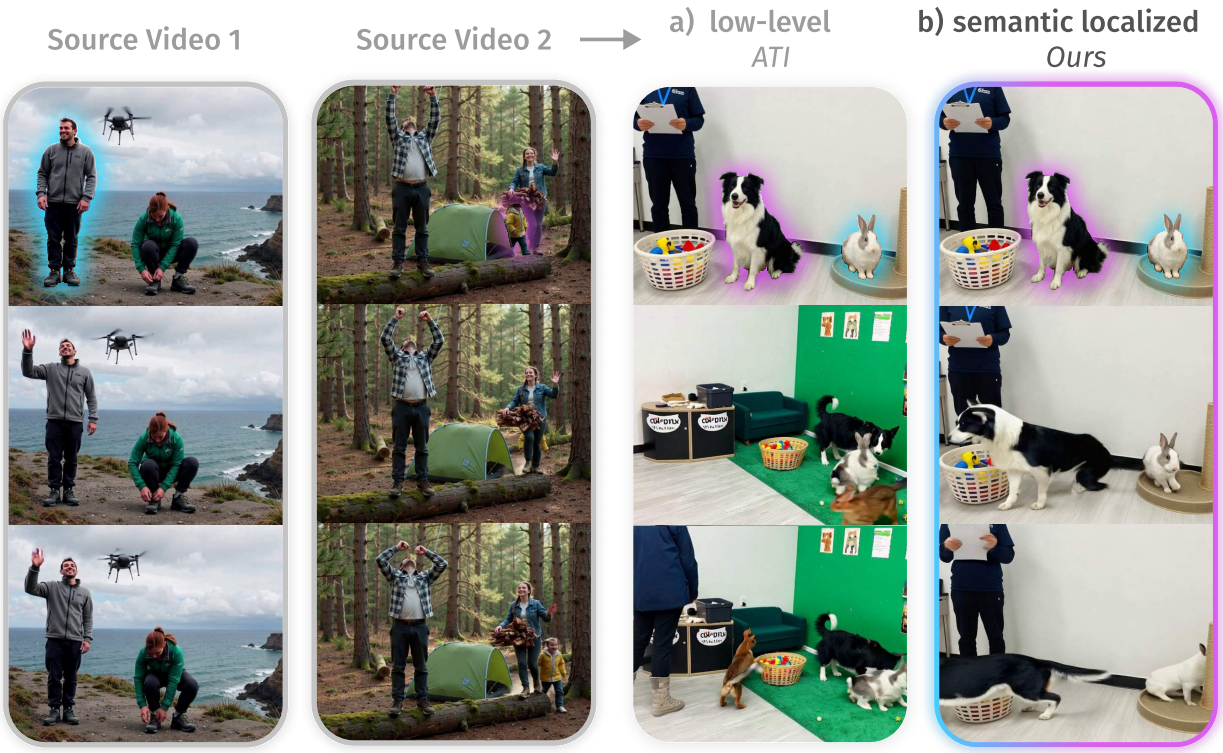


Figure C.13: Composite Motion Transfer. (a) creates degenerate camera motion and does not transfer the correct motions, while **(b)** transfers the correct motion to the right subjects.

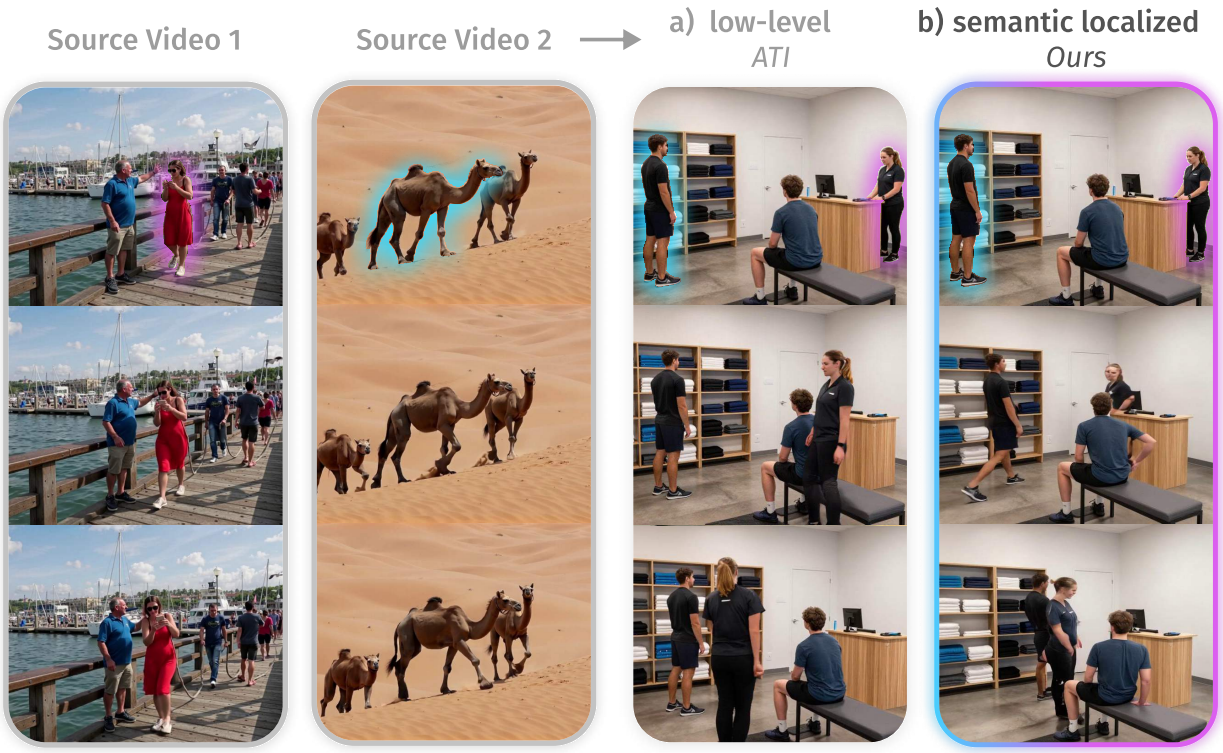


Figure C.14: Composite Motion Transfer. (a) does not transfer the correct motions, while **(b)** transfers the correct motion to the right subjects.



Figure C.15: Text-prompted generation. Top: Driving video providing the articulated motion representation. **Bottom:** Generated video conditioned on the extracted motion and a text prompt, without using a target start frame.