

What Moves?

Localized Motion Representations for Compositional Scene Control

Frank Fundel*, Malek Ben Alaya*, Thomas Ressler-Antal*,
Stefan Andreas Baumann, and Björn Ommer

CompVis @ LMU Munich, Germany
Munich Center for Machine Learning (MCML)

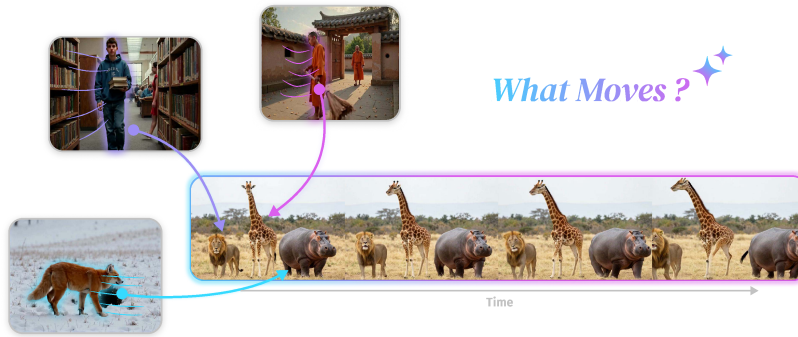


Fig. 1: Scene composition using localized motion representations. Motion is extracted from multiple source videos and attached to target objects, enabling compositional transfer of distinct motions within a single scene.

Abstract. 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>.

* Equal contribution.

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, 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 [35], 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 [12, 28, 35, 37–39], 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 en-

ables 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 [15, 36, 41], many works extend pretrained image backbones with temporal modules to generate coherent video sequences [6, 13]. 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, 14, 21, 45, 55]. 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, 19, 46–48, 57, 58]. 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, 10, 40, 49, 59, 60]. Other approaches guide generation directly using motion signals such as optical flow, temporally warped noise, or related geometric cues [5, 22, 62]. 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 [20, 24, 32, 38, 50, 56, 61]. More recent work explores motion-guided editing paradigms that modify subject appearance while preserving source motion [25, 26]. Other approaches focus on efficiency or representation abstraction, such as training-free motion transfer formulations or depth-aware motion embeddings [28, 37]. 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 [17, 35]. 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.

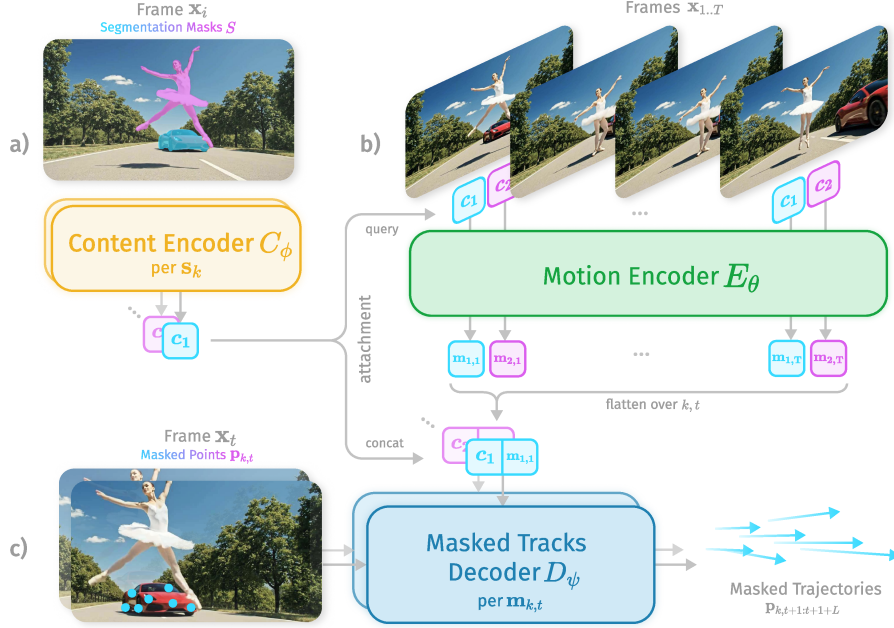


Fig. 2: Method Overview. Given a video frame with corresponding object masks, the content encoder produces static region tokens for appearance grounding (a). The motion encoder (b) extracts persistent, region-specific motion tokens while preserving global context. The masked flow decoder (c) predicts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

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.

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., $C_\phi : (\mathbf{x}_i, \mathbf{s}_k) \rightarrow \mathbf{c}_k$. 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 $\mathbf{c}_k$, 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

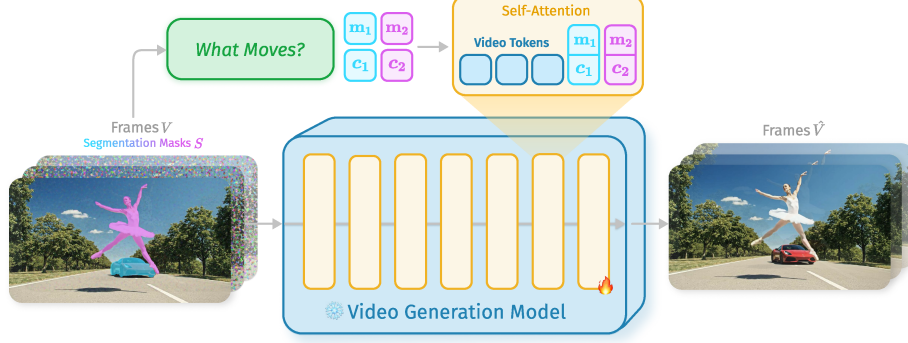


Fig. 3: Video Model Conditioning. With our localized representations, we fine-tune a large video generation model. For this, we append to each temporal self-attention layer of the model, the respective motion embeddings $\mathbf{m}_{1:K,t}$ and content embeddings $\mathbf{c}_{1:K}$ for all regions K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

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 ,

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

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:

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

The backbone is fine-tuned using LoRA [16] 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 [31] frame embedding stage. The masked flow decoder is also implemented as a Transformer[86M

parameters]. We jointly train the encoder and decoder using AdamW [27] for 600k steps with a batch size of 128 on open-world videos from OpenVid-1M [30] and an internally collected dataset, enabling the model to learn general-purpose motion representations. For motion transfer tasks, we utilize CogVideoX-5B [54] 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 [38], originally gathered from DAVIS [33] and VOS [52], and extract subject masks using the promptable segmentation model SAM2 [34]. For each source video, we generate multiple target initial frames using Z-Image Turbo [43].

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

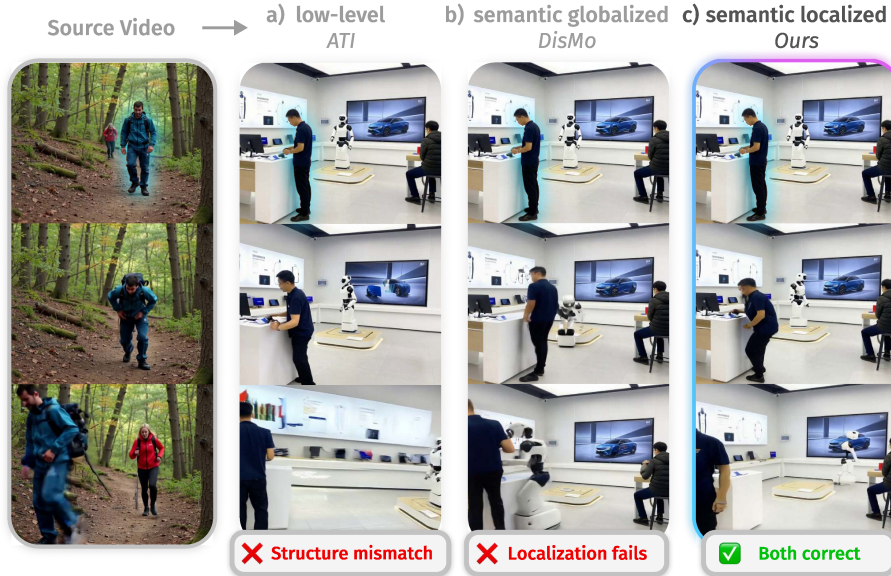


Fig. 4: Motion Transfer Paradigms. Existing methods are typically either localized or semantic. (a) Low-level approaches transfer exact motion and implicitly assume identical structure; while localized, they do not generalize across different structures. (b) Semantic motion transfer methods generalize 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.

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 [35], 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

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.

	Model	video model properties		motion transfer properties		
		Temporal Consistency [↑]	Source Similarity [↑]	In-Region Motion Fidelity [↑]	Out-of-Region Motion Leakage [↓]	Motion Selectivity [↑]
low-level	ATI [1] (Wan 2.1-14B [45])	0.9725	0.5930	0.8111	0.6108	0.2003
	WanMove [7] (Wan 2.1-14B)	<u>0.9738</u>	0.5949	<u>0.7438</u>	<u>0.5608</u>	0.1830
	Tora [59] (CogVideoX-5B [54])	0.9758	0.5807	0.7021	0.4738	<u>0.2283</u>
semi-antic	DisMo [35] (CogVideoX-5B)	<u>0.9657</u>	0.5947	<u>0.6981</u>	<u>0.5604</u>	0.1377
	Ours (CogVideoX-5B)	0.9690	<u>0.5888</u>	0.7462	0.5158	0.2304

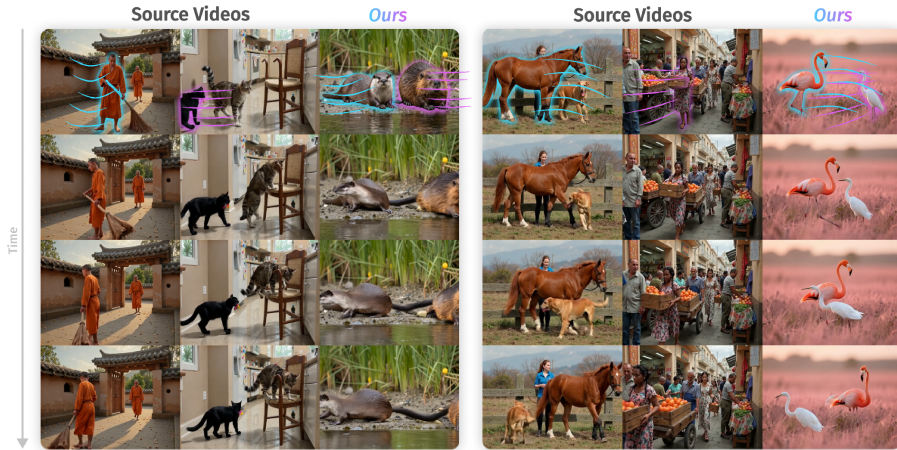


Fig. 5: Composite Qualitative Examples. We provide several examples, showcasing our capability to compose scenes from localized motion embeddings of multiple source videos. Our model correctly isolates and attaches the motion of multiple subjects into one coherent scene.

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 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 [51], 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_{RGB}**),

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 [45]	14B	1.96	4.90	5.23
WanMove [7]	WAN2.1	14B	14.05	14.38	14.38
Tora [59]	CogVideoX [54]	5B	13.73	17.32	15.69
DisMo [35]	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.

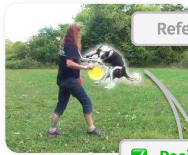
Model	Temporal Consistency [↑]	In-Region Motion Fidelity [↑]	Out-of-Region Motion Leakage [↓]	Motion Selectivity [↑]
ATI [1] (Wan 2.1-14B [45])	0.9634	0.7148	<u>0.3857</u>	<u>0.3291</u>
WanMove [7] (Wan 2.1-14B)	<u>0.9792</u>	<u>0.7059</u>	0.4110	0.2949
Tora [59] (CogVideoX-5B [54])	0.9766	0.6520	0.3904	0.2616
Ours (CogVideoX-5B)	0.9831	0.6718	0.2604	0.4114

and one where we do not apply localized querying at all (**Ours_{GLOBAL}**). Results are summarized in Tab. 4.

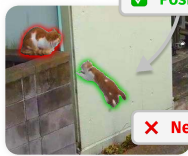
Context is Important for Localized Motion Understanding. Image-based representations such as DINOv2 [31] perform poorly on this task, highlighting the limitations of appearance features for localized action recognition. Video representations such as V-JEPA2 [2] improve performance but remain constrained by their lack of explicit motion modeling. Motion-focused approaches such as DisMo [35] 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_{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_{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 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_{RGB}**) or removing localization by querying the entire frame (**Ours_{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↑	
<i>Image encoders</i>		
DINOv2 _{RGB} [31]	10.9	17.4
DINOv2 _{LATENT}	6.0	10.2
<i>Video encoders</i>		
V-JEPA2 _{RGB} [2]	14.6	23.8
V-JEPA2 _{LATENT}	13.9	21.6
<i>Global motion encoders</i>		
DINOv2 _{RGB} + SemanticMoments [18]	12.4	18.8
DINOv2 _{LATENT} + SemanticMoments	5.5	9.5
V-JEPA2 _{RGB} + SemanticMoments	14.3	23.4
V-JEPA2 _{LATENT} + SemanticMoments	11.9	18.8
DisMo _{RGB} [35]	22.4	33.9
Ours_{RGB}	20.0	31.1
Ours_{GLOBAL}	23.4	35.1
Ours	29.9	42.0



Reference



✓ Positive



✗ Negative

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 [53], which measures action recognition robustness under low-light conditions; IARD [42], which evaluates identity-invariant motion understanding by minimizing reliance on appearance cues; Jester [29], a large-scale hand-gesture dataset that requires sensitivity to fine-grained local motions; Something-Something V2 [11], which focuses on subtle human-object interactions and temporal reasoning; and Diving48 [23], a challenging fine-grained action recognition benchmark designed to reduce scene and object bias, requiring recognition based primarily

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 $\uparrow$	Jester $\uparrow$	Diving48 $\uparrow$	ARID $\uparrow$	IARD $\uparrow$
DINOv2 [31]	10.3	23.2	9.39	14.5	76.1
VideoMAE [44]	7.1	20.1	-	17.3	73.4
V-JEPA2 [2]	22.2	40.8	11.1	28.0	87.6
SemanticMoments _{DINO} [18]	11.4	37.3	9.85	22.0	62.7
SemanticMoments _{V-JEPA2} [18]	31.5	52.2	12.7	38.3	<u>92.5</u>
DisMo [35]	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 $\uparrow$
DINOv2	44.4
VideoMAE	79.2
V-JEPA2	74.4
SemanticMoments _{DINO}	86.4
SemanticMoments _{V-JEPA2}	84.4
DisMo	90.0
Ours	<u>89.6</u>

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.

We evaluate on the synthetic split of the *SemanticMoments* benchmark [18], 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 [31], V-JEPA2 [2], and DisMo [35], 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.

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 BMW 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 Konrad 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. Angtian, W., Haibin, H., Zhiyuan, F., Yiding, Y., Chongyang, M.: ATI: Any trajectory instruction for controllable video generation. arXiv preprint **arXiv:2505.22944** (2025)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), <https://openai.com/research/video-generation-models-as-world-simulators>
5. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., Ryoo, M., Debevec, P., Yu, N.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise (2025), <https://arxiv.org/abs/2501.08331>
6. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7310–7320 (June 2024)
7. Chu, R., He, Y., Chen, Z., Zhang, S., Xu, X., Xia, B., Wang, D., Yi, H., Liu, X., Zhao, H., Liu, Y., Zhang, Y., Yang, Y.: Wan-move: Motion-controllable video generation via latent trajectory guidance (2025), <https://arxiv.org/abs/2512.08765>
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=YicbFdNTTy>
9. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7346–7356 (2023)
10. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Doersch, C., Aytar, Y., Rubinstein, M., Sun, C., Wang, O., Owens, A., Sun, D.: Motion prompting: Controlling video generation with motion trajectories. arXiv preprint arXiv:2412.02700 (2024)
11. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: CVPR. pp. 5842–5850 (2017)
12. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., 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. pp. 1–12 (2025)
13. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning (2024), <https://arxiv.org/abs/2307.04725>

14. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), <https://arxiv.org/abs/2106.09685>
17. Huberman, S., Goldberg, K., Patashnik, O., Benaim, S., Mokady, R.: Semanticmoments: Training-free motion similarity via third moment features (2026), <https://arxiv.org/abs/2602.09146>
18. Huberman, S., Goldberg, K., Patashnik, O., Benaim, S., Mokady, R.: Semanticmoments: Training-free motion similarity via third moment features. arXiv preprint arXiv:2602.09146 (2026)
19. Jain, Y., Nasery, A., Vineet, V., Behl, H.: Peekaboo: Interactive video generation via masked-diffusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8079–8088 (2024)
20. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In: *CVPR*. pp. 9212–9221 (2024)
21. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
22. Koroglu, M., Caselles-Dupré, H., Jeanneret, G., Cord, M.: Onlyflow: Optical flow based motion conditioning for video diffusion models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 6226–6236 (2025)
23. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 513–528 (2018)
24. Ling, P., Bu, J., Zhang, P., Dong, X., Zang, Y., Wu, T., Chen, H., Wang, J., Jin, Y.: Motionclone: Training-free motion cloning for controllable video generation. arXiv preprint arXiv:2406.05338 (2024)
25. Liu, P., Wang, J., Shen, Y., Mo, S., Qi, C., Ma, Y.: Multimotion: Multi subject video motion transfer via video diffusion transformer (2025), <https://arxiv.org/abs/2512.07500>
26. Liu, Y., Sun, Y., Xing, Z., Gao, J., Chen, K., Pei, W.: Motionshot: Adaptive motion transfer across arbitrary objects for text-to-video generation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11861–11871 (2025)
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), <https://arxiv.org/abs/1711.05101>
28. Ma, Y., Wang, Z., Ren, T., Zheng, M., Liu, H., Guo, J., Fong, M., Xue, Y., Zhao, Z., Schindler, K., et al.: Fastvmt: Eliminating redundancy in video motion transfer. arXiv preprint arXiv:2602.05551 (2026)
29. Materzynska, J., Berger, G., Bax, I., Memisevic, R.: The jester dataset: A large-scale video dataset of human gestures. In: *Proceedings of the IEEE/CVF international conference on computer vision workshops*. pp. 0–0 (2019)
30. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In: *ICLR* (2025), <https://openreview.net/forum?id=j7kdXSrISM>

31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (Apr 2023). <https://doi.org/10.48550/arXiv.2304.07193>, <http://arxiv.org/abs/2304.07193>, arXiv:2304.07193 [cs]
32. Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers (2025), <https://arxiv.org/abs/2412.07776>
33. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Gool, L.V.: The 2017 davis challenge on video object segmentation (2018), <https://arxiv.org/abs/1704.00675>
34. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024), <https://arxiv.org/abs/2408.00714>
35. Ressler-Antal, T., Fundel, F., Alaya, M.B., Baumann, S.A., Krause, F., Gui, M., Ommer, B.: Dismo: Disentangled motion representations for open-world motion transfer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), <https://openreview.net/forum?id=jneVld5iZw>
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
37. Sheng, M., Gu, Z., Li, P., Lin, C., Guo, H.X., Chen, Y.C., Liu, Y.: Flexam: Flexible appearance-motion decomposition for versatile video generation control. arXiv preprint arXiv:2602.13185 (2026)
38. Shi, Q., Wu, J., Bai, J., Zhang, J., Qi, L., Li, X., Tong, Y.: Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. arXiv preprint arXiv:2503.17350 (2025)
39. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
40. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Schechtman, E., Huang, X.: Motionstream: Real-time video generation with interactive motion controls (2025), <https://arxiv.org/abs/2511.01266>
41. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
42. Tacchetti, A., Isik, L., Poggio, T.: Invariant Action Recognition Dataset (2019). <https://doi.org/10.7910/DVN/DMTOPG>, <https://doi.org/10.7910/DVN/DMTOPG>
43. Team, I., Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., Li, Z., Li, Z.Y., Liu, D., Liu, D., Shi, J., Wu, Q., Yu, F., Zhang, C., Zhang, S., Zhou, S.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer (2025), <https://arxiv.org/abs/2511.22699>
44. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* **35**, 10078–10093 (2022)

45. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), <https://arxiv.org/abs/2503.20314>
46. Wang, J., Zhang, Y., Zou, J., Zeng, Y., Wei, G., Yuan, L., Li, H.: Boximator: Generating rich and controllable motions for video synthesis. arXiv preprint arXiv:2402.01566 (2024)
47. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
48. Wu, W., Li, Z., Gu, Y., Zhao, R., He, Y., Zhang, D.J., Shou, M.Z., Li, Y., Gao, T., Zhang, D.: Draganything: Motion control for anything using entity representation. In: European Conference on Computer Vision. pp. 331–348. Springer (2024)
49. Xiao, Z., Ouyang, W., Zhou, Y., Yang, S., Yang, L., Si, J., Pan, X.: Trajectory attention for fine-grained video motion control. arXiv preprint arXiv:2411.19324 (2024)
50. Xiao, Z., Zhou, Y., Yang, S., Pan, X.: Video diffusion models are training-free motion interpreter and controller. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), <https://openreview.net/forum?id=ZvQ4Bn75kN>
51. Xu, C., Hsieh, S.H., Xiong, C., Corso, J.J.: Can humans fly? action understanding with multiple classes of actors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2264–2273 (2015)
52. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark (2018), <https://arxiv.org/abs/1809.03327>
53. Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J., See, S.: 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. pp. 70–84. Springer (2021)
54. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer (2025), <https://arxiv.org/abs/2408.06072>
55. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
56. Yatim, D., Fridman, R., Bar-Tal, O., Kasten, Y., Dekel, T.: Space-time diffusion features for zero-shot text-driven motion transfer. In: CVPR. pp. 8466–8476 (2024)
57. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
58. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)

59. Zhang, Z., Liao, J., Li, M., Dai, Z., Qiu, B., Zhu, S., Qin, L., Wang, W.: Tora: Trajectory-oriented diffusion transformer for video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2063–2073 (2025)
60. Zhang, Z., Liao, J., Meng, X., Qin, L., Wang, W.: Tora2: Motion and appearance customized diffusion transformer for multi-entity video generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9434–9443 (2025)
61. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J.W., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. In: European Conference on Computer Vision. pp. 273–290. Springer (2024)
62. Zhou, H., Wang, C., Nie, R., Liu, J., Yu, D., Yu, Q., Wang, C.: Trackgo: A flexible and efficient method for controllable video generation (2025), <https://arxiv.org/abs/2408.11475>