Controlling Motion Transfer in Diffusion Transformers via Attention Heads

Sunyoung Jung[1*], Jiwoo Park[1,2*], Yoonseok Choi[1], Kyobin Choo[1], Ming-Hsuan Yang[3], and Seong Jae Hwang[1†]

[1]Yonsei University [2]LG Electronics [3]University of California, Merced

Project page: `https://sunyj-hxppy.github.io/halo`

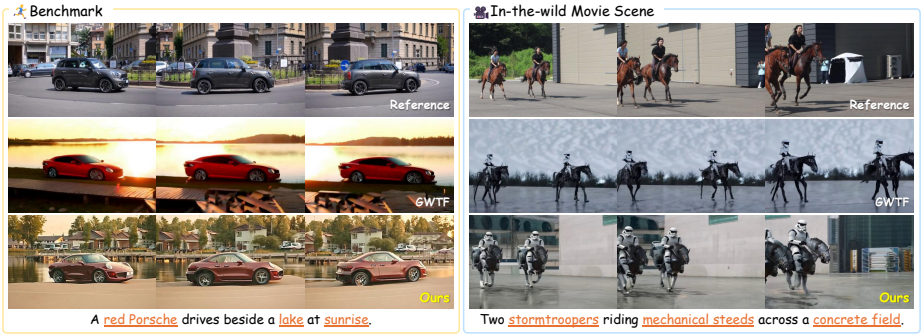

Fig. 1: Overview. We present *HALO*, a head-aware controllable motion transfer framework for video Diffusion Transformers, which identifies motion- and structure-specialized attention heads within the model. Leveraging these findings, *HALO* generates videos that follow the target prompt while remaining motion- and structure-aligned with reference videos, achieving accurate motion transfer.

Abstract. Diffusion Transformers (DiTs) have advanced video generation with high-quality, temporally coherent results. However, extending them to motion transfer, which requires following reference motion while aligning with a target prompt, remains challenging due to limited understanding of motion and structure representations within DiTs. We analyze video DiTs at the attention-head level and identify distinct heads specialized for motion and spatial structure. Based on this insight, we propose a head-aware controllable motion transfer framework that requires no parameter updates. Our method refines motion cues from motion-specialized heads via semantic correspondence guidance and preserves structure through selective feature injection. This head-level control not only enables accurate motion transfer but also provides an interpretable foundation for controllable video generation with DiTs.

Keywords: Motion Transfer · Diffusion Transformers · Attention Heads

1 Introduction

Motion transfer in video generation synthesizes a video that follows the motion of a reference video while adhering to a target prompt. The primary objectives

* Equal contribution. † Corresponding author.

are (1) motion fidelity, ensuring temporal adherence to the reference motion, and (2) structural alignment, maintaining spatial layout of the reference [13,44]. Achieving these goals requires modeling of spatio-temporal dependencies, an area in which recent video Diffusion Transformers (DiTs) [23, 37, 42, 43] have shown strong capability. Given their ability to capture spatial structure and temporal dynamics, DiTs have become a natural choice for motion transfer [7, 13, 32].

Existing DiT-based motion transfer approaches, such as noise warping [7], rotary positional embedding manipulation [13], and cross-frame attention optimization [32], offer varying degrees of motion controllability. However, these methods focus on manipulating motion representations without an understanding of how motion and structure are encoded within DiTs. This lack of understanding is due to the distributed functionality of DiTs, which makes their internal mechanisms challenging to analyze [3, 8, 11]. Consequently, generated videos often exhibit seemingly plausible motion yet inaccurate trajectories or misaligned object structures relative to the reference video. As illustrated in Fig. 1, Go-with-the-flow (GWTF) [7], a state-of-the-art method, exhibits motion deviations, including a car moving straight instead of turning and two stormtroopers with misaligned positions compared to the reference.

Therefore, we conduct a detailed analysis of video DiTs, focusing on the attention heads to answer the fundamental question: *How are motion and structural cues internally encoded in video DiTs?* To identify the heads specialized in modeling motion, we introduce the first head-level analysis based on displacement maps. The displacement map [32] encodes motion as patch-wise coordinate differences between frames, making it effective for analyzing motion properties of heads. For structural cues, we compute the visual token attention-map entropy, which represents the uncertainty in patch dependency. By utilizing the entropy, we select heads that reliably capture structural information.

Our analysis identifies two distinct attention head subsets within video DiTs: (1) *Motion-Specific Heads.* A subset of heads exhibits strong patch correspondences across frames, capturing temporally coherent motion cues in the displacement maps. These heads present clear displacement maps that accurately reflect the motion, demonstrating their superior capability to represent coherent motion flow. (2) *Structure-Specialized Heads.* Another subset of heads encodes structural information. The attention maps of the heads show low entropy, characterized by sharply diagonal patterns. This attention pattern yields attention features with concentrated structural information, confirming that attention map entropy serves as an indicator of the structural content.

Building on these findings, we propose *HALO*, a head-driven semantic-structu ral motion transfer framework that simultaneously enforces motion fidelity and structural alignment. By leveraging the properties of attention heads in video DiTs, our approach first constructs inter-frame displacement maps using only motion-specific heads to guide motion optimization. These displacement maps capture the motion dynamics but lack semantic information. This often leads to transferring motion into semantically irrelevant regions, causing inconsistent or reversed object motion, as shown in Fig. 2(a). To overcome this issue, we

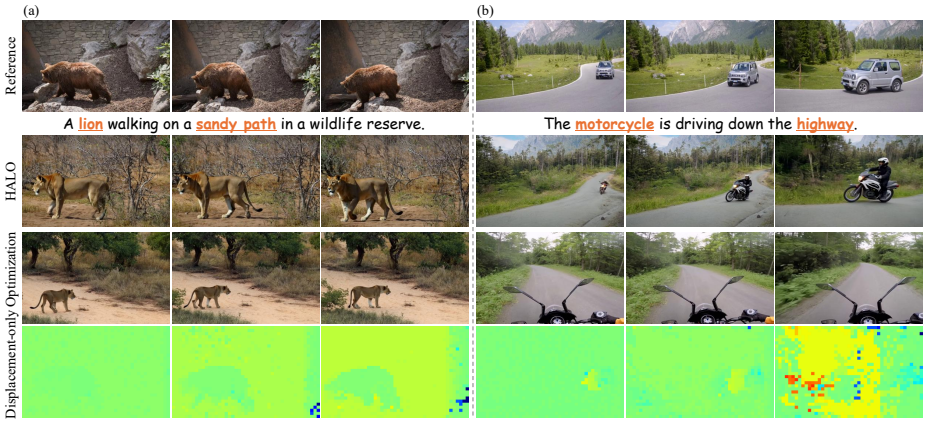

Fig. 2: Limitations of displacement-based motion transfer. Comparison between *HALO* and displacement-only optimization [32]. (a) Lack of semantic alignment causes motion errors. (b) Missing structural preservation leads to spatial misalignment. *HALO* ensures consistent motion and spatial fidelity.

introduce semantic guidance modules that align motion cues with semantic similarities derived from diffusion features, known to encode rich object-level semantics [35, 38]. This refinement produces semantically aligned displacement maps, enabling temporally consistent and semantically coherent motion transfer.

While the refined displacement map provides motion flow information, it remains insufficient for preserving the spatial layout of the reference video. Consequently, the generated results often exhibit structural misalignment with the reference, as illustrated in Fig. 2(b). To overcome this limitation, we leverage our second insight that low-entropy attention heads primarily encode structural information. Specifically, we introduce selective structural feature injection by incorporating attention features from the reference video through these low-entropy heads. This head selection reinforces structural information while suppressing noisy and diffuse cues, yielding motion transfer that is structurally aligned with the reference.

The main contributions of this work are:

- We present a **head-level functional analysis** of video DiTs, revealing motion-specific and structure-specialized attention heads and validating them as **control primitives** for controllable video generation.
- We propose ***HALO***, a *head-aware controllable motion transfer* framework that enhances motion fidelity through **semantic-aware displacement optimization** and preserves spatial consistency via **selective feature injection** from structurally informative heads.
- We perform extensive evaluations on standard motion transfer benchmarks and our Movie Scene Dataset, demonstrating that it achieves **better motion coherence and structural alignment**.

2 Related Work

Text-to-Video Generation. Early Text-to-Video (T2V) generation was driven by U-Net-based diffusion models [9, 21, 26], which extended image diffusion with temporal attention or 3D convolutions to capture spatio-temporal dynamics [4, 6, 14]. More recently, DiTs have emerged as the dominant paradigm [10, 15, 23, 37, 43], significantly improving temporal consistency and visual quality. Motivated by these advances, we build our framework upon a video DiT backbone.

Motion Transfer. Motion transfer aims to generate a target video by extracting and reflecting only the motion information from a reference video [12, 30, 34, 44]. The main challenge of this task lies in effectively decoupling the motion and appearance information from the reference [19, 41].

Early studies have addressed this challenge by extracting motion embedding from the temporal modules of U-Net [24, 39, 46]. Recently, the strong generative capability of DiTs has motivated their adoption in the motion transfer field. For instance, GWTF [7] explicitly warps noise to manipulate the motion, while RoPECraft [13] handles motion by warping the RoPE embeddings using optical flow. Furthermore, DiTFlow [32] derives displacement from cross-frame attention to guide patch movement. While these prior works have explored various ways to handle motion information in DiTs, there has been a lack of direct analysis of how motion and structure are encoded within DiTs' internal representations.

To bridge this analytical gap, our work analyzes DiT internal representations to identify motion-specific and structure-specialized heads. By leveraging these heads, we achieve training-free motion transfer and extend our approach to controllable video generation.

Attention in Video DiTs. U-Net-based video diffusion models [9, 21] explicitly separate spatial and temporal attention, enabling motion extraction from temporal modules [19, 24]. In contrast, DiTs [37, 43] adopt unified attention that jointly models spatial and temporal information. While this design enhances expressiveness, it obscures the distinction between motion and structure, complicating motion-specific control [25, 32]. In contrast, we posit that this motion information is independently encoded within the unified structure of attention in the DiTs. Consequently, we focus on analyzing this unified attention structure at the level of individual attention heads to identify the distinct encoding of motion and structure in terms of motion transfer.

3 Analyzing Attention Heads of Video DiTs

We focus on understanding how motion and structural information are represented within video DiTs [43]. Here, we present a head-level analysis of attention heads to uncover their distinct functional roles in modeling motion and spatial structure, offering a mechanistic perspective overlooked in prior studies [1, 22, 40].

Analysis 1: Motion-Specific Heads. We aim to identify the motion-specific properties encoded within individual attention heads. Building on SparseGen [40], which identifies temporal and spatial head patterns for efficient video generation

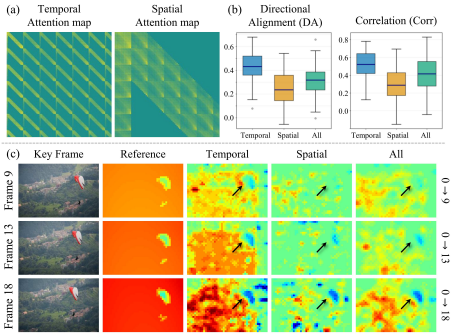

Fig. 3: Head Configuration Comparison. (a) Attention maps show distinct patterns: *temporal heads* capture cross-frame diagonals, while *spatial heads* maintain intra-frame locality. (b) Quantitative evaluation using directional alignment and correlation shows that temporal heads align more closely with reference motion. (c) Displacement maps further confirm temporal heads more accurately capture motion flow.

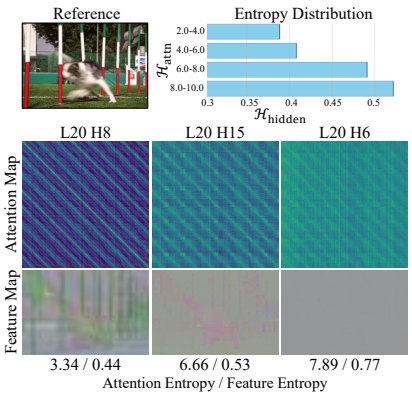

Fig. 4: Relationship between attention maps and structural cues in hidden features. Heads with lower attention entropy show lower feature entropy, indicating stronger structural fidelity. Example: L20 H8 denotes the 8th attention head in the 20th layer.

(see Fig. 3(a)), we first classify heads based on their similarity to pattern masks (e.g., cross-frame diagonal for temporal, sparse localized patterns for spatial). Based on this, we construct head-specific displacement maps from cross-frame attention [28,32]. Specifically, cross-frame attention is computed using the queries Q and keys K derived from the latent features $z_f \in \mathbb{R}^{head \times N \times d}$ as:

$$\mathbf{M}^h_{f,f'} = \text{softmax}\left(\frac{Q^h_f \cdot K^{h\top}_{f'}}{\sqrt{d}}\right), \tag{1}$$

where h denotes the attention head, $N = H \times W$ is the number of spatial tokens, and d is the feature dimension. Here, f and f' represent the frame indices. We then derive a head-specific displacement map $\mathcal{D}^h_{f,f'} \in \mathbb{R}^{H \times W \times 2}$, which captures the patch-wise motion between frames (f, f'), defined as:

$$\mathrm{I}^h_{f,f'} = \underset{f'}{\text{argmax}}\left(\mathbf{M}^h_{f,f'}\right), \quad \mathcal{D}^h_{f,f'}(i) = \mathbf{g}\big(\mathrm{I}^h_{f,f'}(i)\big) - \mathbf{g}(i), \tag{2}$$

where $\mathbf{g}(\cdot)$ converts a 1D patch index into 2D spatial coordinates and i denotes the patch index.

Using displacement maps, we analyze cross-frame motion patterns across attention heads to quantify motion capture effectiveness, as shown in Fig. 3(c). Temporal heads capture motion variations more accurately than spatial or combined heads, showing stronger alignment with object motion (blue) and background movement (red). To validate these observations, we measure the similarity between predicted displacement maps and ground-truth optical flow [36]

through two quantitative metrics: (1) Directional Alignment (DA), calculating cosine similarity between motion directions, and (2) Correlation (Corr), using Pearson correlation for global pattern consistency.

As illustrated in Fig. 3(b), temporal heads consistently achieve higher DA and Corr scores. These results confirm that while spatial heads attend to intra-frame patch relationships, temporal heads specialize in cross-frame dependencies, effectively encoding motion-related information within video DiTs.

Analysis 2: Structure-Specialized Heads. To identify attention heads that encode structural information, we analyze the visual token attention maps A^h corresponding to each head. Our analysis reveals distinct patterns across different heads: Heads exhibiting well-defined diagonal alignment patterns with low entropy (e.g., L20 H8) stand in contrast to those with irregular or diffuse attention distributions with high entropy (e.g., L20 H6), as shown in Fig. 4. Since the entropy serves to quantify the uncertainty in capturing the relationships between visual tokens, low entropy implies the presence of structural information.

To validate this, we compute the attention entropy [29] and the spatial entropy of corresponding attention features [5,20,31]. Given a head h with attention map A^h, the attention entropy $\mathcal{H}^h_{\text{attn}}$ is defined as $\mathcal{H}^h_{\text{attn}} = -\sum_{k=1}^{N} A^h_k \log A^h_k$, where lower values indicate a more concentrated attention distribution. To capture spatial properties in attention features, we compute the spatial entropy of PCA-transformed feature embeddings.

As illustrated in Fig. 4, lower entropy heads exhibit clear diagonal attention patterns and yield feature maps maintaining structural layouts. Fig. 4 also provides quantitative evidence of a strong correlation between lower attention entropy $\mathcal{H}_{\text{attn}}$ and lower hidden feature spatial entropy $\mathcal{H}_{\text{hidden}}$. Therefore, these diagonal attention patterns indicate strong spatial patch correlations, validating that such heads are specialized in encoding structural information.

4 Method

We introduce *HALO*, a head-aware controllable motion transfer framework that jointly achieves *motion fidelity* and *structural alignment* with a reference video (see Fig. 5(a)). Building on Sec. 3, we propose (i) semantic-aware motion optimization (Sec. 4.1) and (ii) structure-guided feature injection (Sec. 4.2), illustrated in Fig. 5(b)–(c).

4.1 Semantic-Aware Motion Guidance

For motion transfer, we construct cross-frame displacement maps within DiTs. Our analysis (Analysis 1, Sec. 3) shows that motion-specific heads $\mathcal{M}$ capture motion more effectively. We therefore aggregate cross-frame attention $\mathbf{M}^h_{f,f'}$ for $h \in \mathcal{M}$, identified via temporal pattern masks to extract a displacement map.

While cross-frame attention provides patch-level correlation cues, it can overemphasize visually similar yet semantically unrelated regions, producing mismatched displacements and degraded motion alignment (e.g., camera-following movements that disregard objects and incorrect object trajectories; see Supp. Sec. B).

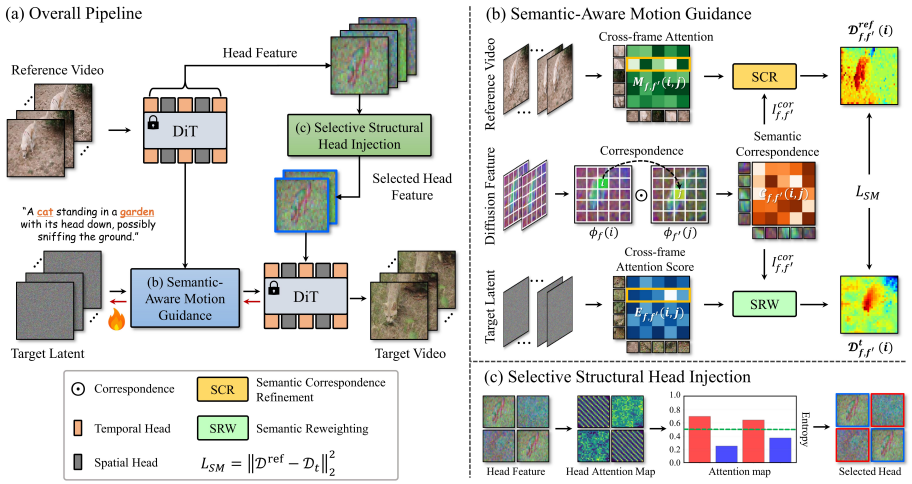

Fig. 5: Overview of *HALO*. (a) From a reference video, we extract displacement maps and head features. Displacements from motion-specific heads guide motion by optimizing the latent representation, while selected head features from the reference are injected to preserve structure during generation. (b) To enhance motion guidance, semantic correspondence derived from diffusion features refines the displacement map, ensuring semantically aligned motion flow. (c) For structural guidance, we inject value features from heads selected via entropy analysis, targeting heads that encode essential spatial information.

To mitigate this, we introduce a semantic-aware motion guidance that refines displacements using semantic similarities from semantically rich diffusion features [16, 35, 38], as shown in Fig. 5(b).

Semantic Correspondence Refinement. To refine the reference displacement map, we first select the top-k attention candidates from cross-frame attention rather than relying on a single maximum, as illustrated in Fig. 6(a). We then construct semantic correspondence $\mathcal{C}$ using pairwise cosine similarity over diffusion features ϕ:

$$\mathcal{C}_{f,f'} = \frac{\phi_f \cdot \phi_{f'}}{\|\phi_f\|_2 \|\phi_{f'}\|_2}, \tag{3}$$

For each patch index, we compute the distances between the most similar semantic indices $\mathrm{I}^{\mathrm{cor}}_{f,f'} \in \mathbb{R}^N$ between frames and the top-k attention candidates. The nearest candidate is then selected as the reference best-match coordinate $\mathrm{I}^{\mathrm{ref}}_{f,f'}$. The reference inter-frame displacement map $\mathcal{D}^{\mathrm{ref}}$ is then constructed from $\mathrm{I}^{\mathrm{ref}}_{f,f'}$ following Eq. 2 (first row in Fig. 7).

Motion Optimization. Given $\mathcal{D}^{\mathrm{ref}}$, we optimize z_T to align generated motion with the reference. We apply Semantic Reweighting (SRW) to adjust cross-frame attention via a correspondence-guided bias derived from $\mathcal{C}$ (Fig. 6(b)). Using $\mathrm{I}^{\mathrm{cor}}$,

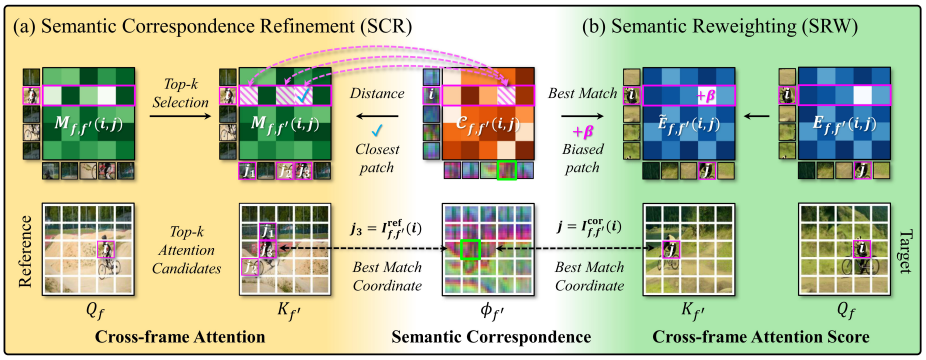

Fig. 6: Details of Semantic Correspondence Refinement (SCR) and Semantic Reweighting (SRW). (a) $\mathrm{I}^{\mathrm{ref}}_{f,f'}$ is obtained by choosing, among top-k attention candidates, the patch closest to the semantic best match $\mathrm{I}^{\mathrm{cor}}_{f,f'}$. (b) SRW manipulates target cross-frame attention by adding a correspondence-based bias at $\mathrm{I}^{\mathrm{cor}}_{f,f'}$.

we form the bias matrix $\mathcal{B}_{f,f'}$:

$$\mathcal{B}_{f,f'}(i,j) = \begin{cases} \beta, & \text{if } j = \mathrm{I}^{\mathrm{cor}}_{f,f'}(i), \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where β controls bias strength. The refined cross-frame attention scores are $\tilde{E}_{f,f'} = E_{f,f'} + \mathcal{B}_{f,f'}$, and refined attention is $\tilde{M}_{f,f'} = \mathrm{softmax}(\tilde{E}_{f,f'})$. We then estimate displacement as an expectation over coordinate differences,

$$\Delta x_{f,f'} = \sum_{i,j}(x_j - x_i)\tilde{M}_{f,f'}(i,j), \quad \Delta y_{f,f'} = \sum_{i,j}(y_j - y_i)\tilde{M}_{f,f'}(i,j), \tag{5}$$

and stack them in $\mathcal{D}_t = [\Delta x_{f,f'}, \Delta y_{f,f'}]$ (second row in Fig. 7). Finally, to optimize the latent for motion alignment, we minimize the semantic motion loss L_{SM}, defined as the L2 distance between the reference displacement map $\mathcal{D}_{\mathrm{ref}}$ and the target displacement map $\mathcal{D}_t$. Incorporating semantic cues into displacement estimation unifies motion alignment with semantic coherence, yielding consistent motion transfer. Notably, this strategy represents the first attempt to incorporate semantic correspondences within motion representations, facilitating high-fidelity, semantically consistent video generation.

4.2 Selective Structural Head Injection

Although displacement maps capture directional motion, they lack sufficient intra-frame structure, which can cause spatial misalignment with the reference (Fig. 2, right). To address this, we propose a structural guidance mechanism for preserving the spatial integrity of the reference. We inject reference value features during generation. Naïvely injecting all head features, however, introduces artifacts such as noise and identity leakage [2] (see Supp. Sec. C.4, Supp. Fig. 6).

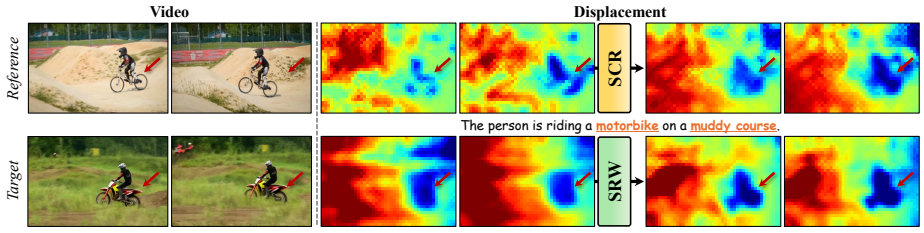

Fig. 7: Effect of SCR and SRW on displacement maps. The refinements improve robustness to fine-grained object motion and better preserve object shape.

We therefore select structurally informative heads via entropy-based analysis (Analysis 2, Sec. 3). Low-entropy heads exhibit strong diagonal attention and produce spatially coherent features; we inject their value features into the corresponding heads, supplying structure-aware guidance without noise or leakage (Fig. 5(c)). This selective injection complements motion optimization and enforces spatial alignment with the reference.

5 Experiments

Implementation Details. Following prior works, we evaluate our approach using a standard motion transfer benchmark and settings [13, 33, 44]. Across all experiments, we adopt the video diffusion transformer CogVideoX [43] as the base model, using 50 denoising steps, 12 optimization steps T_{opt}, and a guidance scale of 7, consistent with common practice. For additional results with the video DiT model, Wan [37], please refer to Supp. Sec. D. The bias strength β for semantic reweighting is fixed to 0.1, and the entropy range τ for selective structure specialized-head injection is set to $\tau < 7$, based on 7 being the median entropy value observed in our analysis (validated in Section 5.3). Hyperparameter configurations (e.g., β, T_{opt}, τ, top-k) and *HALO* optimization algorithm are detailed in Supp. Sec. C.1 and C.2, respectively.

Baselines. We compare against representative motion-transfer approaches, including U-Net-based methods (MOFT [41], ConMO [12], and MotionClone [24]) and recent DiT-based methods (DiTFlow [32], RoPECraft [13], and GWTF [7]).

Movie Scene Dataset. Motion transfer has emerged as a pivotal task in video production and digital cinematography, where it is increasingly demanded to replicate sophisticated film-style shot compositions and intricate post-production effects [27]. However, existing benchmarks [24,33] are largely derived from generic internet videos (e.g., DAVIS/WebVid), which do not fully reflect these professional requirements. Therefore, we curate a Movie Scene Dataset specifically for the motion transfer task (see Supp. Sec. E for details). Complementing existing benchmarks, this dataset provides a production-oriented evaluation setting to assess whether methods can achieve faithful motion transfer under real-world cinematic conditions.

Fig. 8: Qualitative comparison between U-Net- and DiT-based baselines and *HALO*.

5.1 Main Results

Qualitative Results. Fig. 8 presents qualitative comparisons with state-of-the-art models. Baseline methods often exhibit camera drift, generation failures, or identity leakage (e.g., generating a motorcycle instead of a horse). They also produce semantically inconsistent motion, such as misaligned orientations between correlated objects (e.g., car vs. airplane). In contrast, our method produces semantically coherent motion aligned with the reference while preserving subject identity and structural integrity. Furthermore, in Fig. 9, *HALO* exhibits remarkable robustness on the Movie Scene Dataset. Even when guided by complex cinematic prompts, *HALO* faithfully preserves the target's identity integrity while following the reference dynamics.

Quantitative Results. We evaluate performance using CLIP score (CLIP) [17] for text–video alignment, Temporal Consistency (TC) [45] for frame coherence, and Motion Fidelity (MF) [44] and FTD [13] for reference-motion alignment. As shown in Table 1, *HALO* achieves notable improvements in CLIP, MF, and FTD, demonstrating that head-level analysis within DiTs enables motion transfer with strong motion fidelity and structural alignment.

Since motion transfer requires following the reference motion while adhering to the target text prompt, CLIP and MF must be jointly evaluated. Prior

Fig. 9: Qualitative results in the movie scene dataset.

works exhibit a trade-off between these metrics, as shown in Table 1. In contrast, our method achieves balanced performance, maintaining high CLIP and MF scores. This indicates that our head-analysis-based formulation transfers reference motion effectively while suppressing irrelevant semantic information. Although *HALO* reports a slightly lower TC than DiTFlow and GWTF in Table 1, TC tends to favor more static outputs. Further analysis of the CLIP–MF balance and the relationship between motion dynamics and temporal consistency is provided in Supp. Sec. F.2. We also report the runtime and peak GPU memory in Supp. Sec. F.3, showing that *HALO* incurs only marginal overhead relative to the displacement optimization framework.

User Study. To ensure an assessment aligned with human preference, we conduct a user study to evaluate *HALO* across three criteria: editing accuracy, temporal consistency, and motion accuracy. Twenty participants rated each result on a 5-point Likert scale (1: lowest, 5: highest), with scores subsequently normalized to a 0–100 range. As reported in Table 4, *HALO* achieves the highest mean score across all criteria, demonstrating a consistent preference among participants.

5.2 Ablation Study

We perform ablation studies under conditions identical to the main experiments, using displacement optimization [32] as the baseline.

Effect of Semantic-Aware Motion Guidance. To evaluate the role of semantic correspondence refinement in displacement computation, we directly compute displacement maps from raw cross-frame attention. As shown in Fig. 10 (green

Table 1: Quantitative results with state-of-the-art motion transfer methods. Best and second results are represented with **bold** and underlined.

Model			CLIP ↑	TC ↑	MF ↑	FTD ↓
U-Net-based	MoFT [41]	NeurIPS'24	30.9	85.8	34.8	23.0
	ConMO [12]	CVPR'25	29.8	85.3	52.0	**17.4**
	MotionClone [24]	ICLR'25	30.4	78.6	55.6	19.8
DiT-based	DiTFlow [32]	CVPR'25	31.0	**89.5**	59.6	23.0
	RoPECraft [13]	NeurIPS'25	30.3	85.9	58.2	19.6
	GWTF [7]	CVPR'25	31.6	88.2	62.5	21.6
	HALO		**31.7**	87.5	**66.2**	19.4

Table 2: Quantitative results in the Movie Scene dataset (methods requiring video masks are excluded).

Model	CLIP ↑	TC ↑	MF ↑
MotionClone [24]	27.1	74.5	47.7
DiTFlow [32]	29.7	**90.4**	48.4
RoPECraft [13]	28.3	88.2	49.1
GWTF [7]	30.2	87.6	46.6
HALO	**30.5**	88.9	**52.6**

Table 3: Ablation results of proposed components on performance. "Semantic" denotes the semantic-aware motion guidance, and "Injection" denotes the selective structural head injection.

Exp.#	Semantic	Injection	CLIP ↑	TC ↑	MF ↑	FTD ↓
1			31.0	89.5	59.6	23.0
2		✓	30.6	88.3	61.8	20.5
3	✓		30.9	88.5	61.3	20.9
4 (*HALO*)	✓	✓	31.7	87.5	66.2	19.4

Table 4: User preference study on motion transfer. The results represent the preference rate across three dimensions.

Model	Edit Acc.	TC	Motion Acc.
DiTFlow	84.1	83.1	69.0
RoPECraft	72.4	75.6	77.6
GWTF	84.0	77.1	72.8
HALO	**86.5**	**87.8**	**93.3**

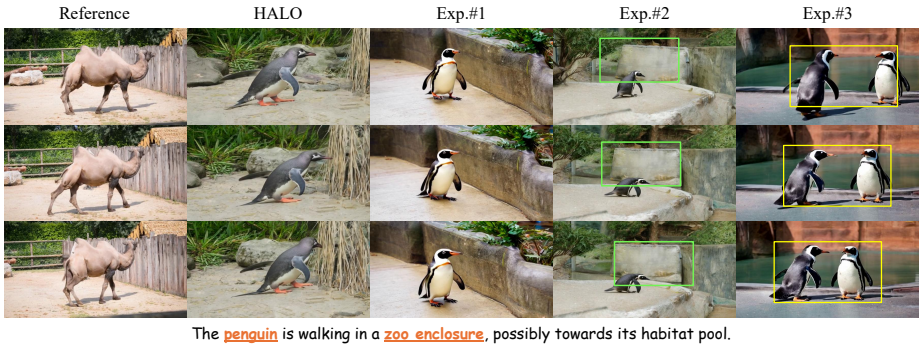

Fig. 10: Qualitative results of our method across ablation studies. Exp.# corresponds to Table 3.

box), the absence of semantic refinement causes motion to the background rather than the actual moving object. This demonstrates that cross-frame attention alone captures visually similar but semantically unrelated regions, leading to motion misalignment. Semantic correspondence refinement is therefore crucial for achieving semantically consistent motion transfer.

Effect of Selective Structure-Specialized Head Injection. Removing selective structure-specialized head injection significantly degrades motion fidelity (Exp.#3 in Table 3). As shown in Fig. 10 (yellow box), this omission results in structural distortions and duplicated objects (e.g., two penguins). These findings indicate that injecting low-entropy structural features helps maintain spatial consistency, making explicit structural guidance essential for accurate motion transfer.

Table 5: Performance of motion-transfer methods across different motion difficulty levels categorized based on the complexity of motion trajectories.

Model	Hard		Medium		Easy	
	CLIP ↑	MF ↑	CLIP ↑	MF ↑	CLIP ↑	MF ↑
MotionClone	26.8	39.0	24.8	37.0	27.8	39.0
ConMo	30.7	51.0	31.0	54.0	28.3	64.0
DiTFlow	32.1	54.0	32.5	67.0	32.1	66.0
RoPECraft	27.1	55.7	24.7	69.9	24.1	59.5
GWTF	**33.6**	56.0	32.7	69.0	30.8	74.0
HALO	33.2	**59.2**	**33.8**	**72.3**	**31.4**	**75.5**

Table 6: Analysis of head configuration and selective injection entropy threshold. (a) Different attention heads. (b) Validation of selective structural head injection entropy range τ.

(a) Attention Head

Head	MF ↑	FTD ↓
All	63.1	21.6
Spatial	53.6	22.7
Motion	66.2	19.4

(b) Entropy Range

τ	MF ↑	FTD ↓
$7 <$	54.1	23.5
< 7	66.2	19.4

5.3 Additional Validations

Complex Motion. Robust motion transfer remains challenging under rapid movements and long-range trajectories. To evaluate the robustness of *HALO* in such settings, we utilize a benchmark stratified into three difficulty levels (Easy/Medium/Hard) based on trajectory complexity. We assess performance using prompt alignment (CLIP) and motion fidelity (MF), while omitting MoFT due to its lack of competitiveness in these scenarios.

As summarized in Table 5, *HALO* consistently achieves superior performance across all levels with minimal performance decay, indicating high stability. This robustness is further evidenced qualitatively in Fig. 11(a); our method handles extreme cases effectively, such as occlusions in a motorcycle jump and the high-velocity dynamics of boxing. Overall, these results validate that our approach remains stable across a wide spectrum of complex motions.

Validation of Motion-Specific Heads. To verify motion-head selection, Table 6a compares motion-specific, spatial, and all-head configurations. Spatial heads alone yield significantly lower MF and FTD, indicating a limited motion-related signal for motion transfer. Although aggregating all heads offers a slight improvement, it still underperforms motion-specific heads. These results suggest that incorporating spatial heads can dilute relevant motion cues; in contrast, isolating motion-specific heads better captures dynamic correlations essential for motion transfer. Qualitative comparisons are included in Supp. Sec. C.5.

To demonstrate the broader applicability of motion-specific heads as a motion-control mechanism, we conduct additional experiments focusing on motion dynamics. Building on Layer Perturbation Guidance [28], we extend this approach to our motion-specific head by introducing head perturbation during inference. Evaluation using VBench metrics [18] on 100 video samples reveals that integrating this method with CogVideoX [43] improves image quality by +14%, aesthetics by +8%, and motion dynamics by +15%. These results suggest that our motion-head analysis provides a lightweight yet effective mechanism to guide motion in DiT-based architectures.

Validation of Structure-Specialized Heads. Table 6b presents results under different entropy thresholds for structure-specialized head selection. Based on

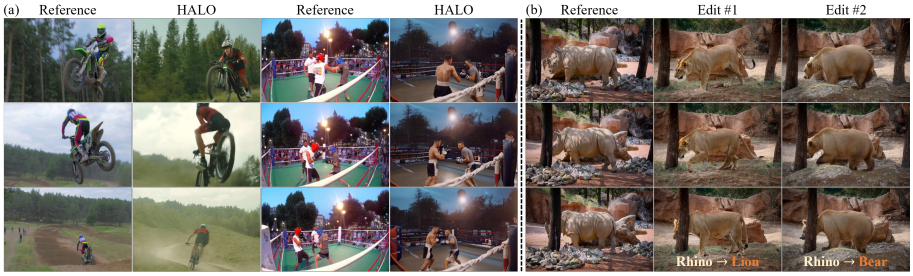

Fig. 11: Qualitative results on complex motion and video editing. (a) Generated samples illustrating the *HALO*'s capability in handling complex motion dynamics. (b) Application to video editing using the structure-specialized head.

our analysis showing a median entropy of 7, we adopt τ=7 as the threshold. Injecting features from high-entropy heads (τ>7) yields lower MF and FTD scores, indicating that such heads contain weaker, more diffuse structural cues. This validates the relationship between attention-map entropy and structural information in the context of motion transfer. Additional experiments over a wider entropy range and qualitative examples are provided in Supp. Sec. C.1.

To further assess the applicability of structure-specialized heads, we extend their utility to video editing, a task where preserving structural integrity is crucial. Specifically, we perform the attention value injection within these selected structure-specialized heads. As shown in Fig. 11(b), our approach successfully modifies subject identity (e.g., rhino→lion/bear) while maintaining the original layout and structural details. These results demonstrate that structure-specialized heads serve as a precise mechanism for structure-preserving video editing, extending their effectiveness beyond motion transfer.

6 Conclusion

In this work, we demonstrate for the first time that video DiTs contain distinct motion-specific and structure-specialized heads responsible for temporal dynamics and spatial organization. Based on this insight, we present *HALO*, a training-free motion transfer framework that leverages these heads for high motion fidelity and structural alignment. Our method integrates semantic correspondences with motion representations and proposes a head-wise injection to enable motion-consistent, structure-aligned transfer. Extensive experiments validate that *HALO* outperforms prior approaches, highlighting the effectiveness of head-level analysis. Furthermore, by introducing our Movie Scene Dataset, we provide a new direction for future research in practical motion transfer.

Future Work. Our findings reveal that video DiTs inherently disentangle motion and structure across attention heads, opening new directions for controllable video generation. Future research may explore explicit control over motion direction and intensity for fine-grained editing.

Acknowledgments

This work was supported in part by the IITP RS-2024-00457882 (AI Research Hub Project), IITP 2020-II201361, NRF RS-2024-00345806, and NRF RS-2023-002620, while the authors are affiliated with the Department of Artificial Intelligence (S.J., J.P, S.J.H.) and the Department of Computer Science (K.C.). This work was supported by the AI Seoul Tech Research Support Program of the Seoul Future Foundation. We would like to express our sincere gratitude to RASCA FX for providing access to the movie dataset used in this work. We also thank the RASCA FX team for their valuable support and collaboration throughout this project.

References

1. Ahn, D., Kang, J., Lee, S., Kim, M., Min, J., Jang, W., Lee, S., Paul, S., Hong, S., Kim, S.: Fine-grained perturbation guidance via attention head selection. arXiv e-prints pp. arXiv–2506 (2025)
2. Atzmon, Y., Gal, R., Tewel, Y., Kasten, Y., Chechik, G.: Motion by queries: Identity-motion trade-offs in text-to-video generation. arXiv preprint arXiv:2412.07750 (2024)
3. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable flow: Vital layers for training-free image editing. In: CVPR. pp. 7877–7888 (2025)
4. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
5. Batty, M.: Spatial entropy. Geographical analysis **6**(1), 1–31 (1974)
6. 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)
7. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., et al.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In: CVPR. pp. 13–23 (2025)
8. Cai, M., Cun, X., Li, X., Liu, W., Zhang, Z., Zhang, Y., Shan, Y., Yue, X.: Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In: CVPR. pp. 7763–7772 (2025)
9. 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: CVPR. pp. 7310–7320 (2024)
10. Chen, S., Xu, M., Ren, J., Cong, Y., He, S., Xie, Y., Sinha, A., Luo, P., Xiang, T., Perez-Rua, J.M.: Gentron: Diffusion transformers for image and video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6441–6451 (2024)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)

12. Gao, J., Yin, Z., Hua, C., Peng, Y., Liang, K., Ma, Z., Guo, J., Liu, Y.: Conmo: Controllable motion disentanglement and recomposition for zero-shot motion transfer. In: CVPR. pp. 7191–7200 (2025)
13. Gokmen, A.B., Ekin, Y., Bilecen, B.B., Dundar, A.: Ropecraft: Training-free motion transfer with trajectory-guided rope optimization on diffusion transformers. arXiv preprint arXiv:2505.13344 (2025)
14. 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. In: ICLR (2024), https://openreview.net/forum?id=Fx2SbBgcte
15. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
16. Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. NeurIPS **36**, 8266–8279 (2023)
17. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
18. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
19. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In: CVPR (2023)
20. Kang, S., Kim, J., Kim, J., Hwang, S.J.: Your large vision-language model only needs a few attention heads for visual grounding. In: CVPR. pp. 9339–9350 (2025)
21. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: ICCV. pp. 15954–15964 (2023)
22. Kim, C., Shin, H., Hong, E., Yoon, H., Arnab, A., Seo, P.H., Hong, S., Kim, S.: Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. arXiv preprint arXiv:2509.18096 (2025)
23. 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)
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, B., Wang, C., Su, T., Ten, H., Huang, J., Guo, K., Jia, K.: Understanding attention mechanism in video diffusion models. arXiv preprint arXiv:2504.12027 (2025)
26. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)
27. Ma, Y., Feng, K., Hu, Z., Wang, X., Wang, Y., Zheng, M., He, X., Zhu, C., Liu, H., He, Y., et al.: Controllable video generation: A survey. arXiv preprint arXiv:2507.16869 (2025)
28. Nam, J., Son, S., Chung, D., Kim, J., Jin, S., Hur, J., Kim, S.: Emergent temporal correspondences from video diffusion transformers. arXiv preprint arXiv:2506.17220 (2025)

29. Pardyl, A., Rypeść, G., Kurzejamski, G., Zieliński, B., Trzciński, T.: Active visual exploration based on attention-map entropy. arXiv preprint arXiv:2303.06457 (2023)
30. Park, G.Y., Jeong, H., Lee, S.W., Ye, J.C.: Spectral motion alignment for video motion transfer using diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6398–6405 (2025)
31. Peruzzo, E., Sangineto, E., Liu, Y., De Nadai, M., Bi, W., Lepri, B., Sebe, N.: Spatial entropy as an inductive bias for vision transformers. Machine Learning **113**(9), 6945–6975 (2024)
32. Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers. In: CVPR. pp. 22911–22921 (2025)
33. Shi, Q., Wu, J., Bai, J., Zhang, J., Qi, L., Tong, Y., Li, X.: Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In: ICCV. pp. 10995–11005 (2025)
34. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: CVPR. pp. 2377–2386 (2019)
35. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. NeurIPS **36**, 1363–1389 (2023)
36. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020)
37. 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. arXiv preprint arXiv:2503.20314 (2025)
38. Wang, J., Ma, Y., Guo, J., Xiao, Y., Huang, G., Li, X.: Cove: Unleashing the diffusion feature correspondence for consistent video editing. NeurIPS **37**, 96541–96565 (2024)
39. Wang, L., Mai, Z., Shen, G., Liang, Y., Tao, X., Wan, P., Zhang, D., Li, Y., Chen, Y.C.: Motion inversion for video customization. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
40. Xi, H., Yang, S., Zhao, Y., Xu, C., Li, M., Li, X., Lin, Y., Cai, H., Zhang, J., Li, D., et al.: Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. arXiv preprint arXiv:2502.01776 (2025)
41. Xiao, Z., Zhou, Y., Yang, S., Pan, X.: Video diffusion models are training-free motion interpreter and controller. NeurIPS **37**, 76115–76138 (2024)
42. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. IEEE Transactions on Visualization and Computer Graphics **31**(2), 1526–1541 (2024)
43. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025), https://openreview.net/forum?id=LQzN6TRFg9
44. 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)

45. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations (2023)
46. 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: ECCV. pp. 273–290. Springer (2024)