QWERTY: Training-Free Motion Control via Query-Warped Video Diffusion Transformers

Kyobin Choo[1], Youngmin Kim[2], Hyunkyung Han[2], Geunrip Park[2], Chanyoung Kim[2], Sunyoung Jung[2], and Seong Jae Hwang[2⋆]

[1] Department of Computer Science, Yonsei University, Seoul, Republic of Korea
[2] Department of Artificial Intelligence, Yonsei University, Seoul, Republic of Korea
{chu, winston1214, hhk, gunlip1210, chanyoung, sunyoungj, seongjae}@yonsei.ac.kr

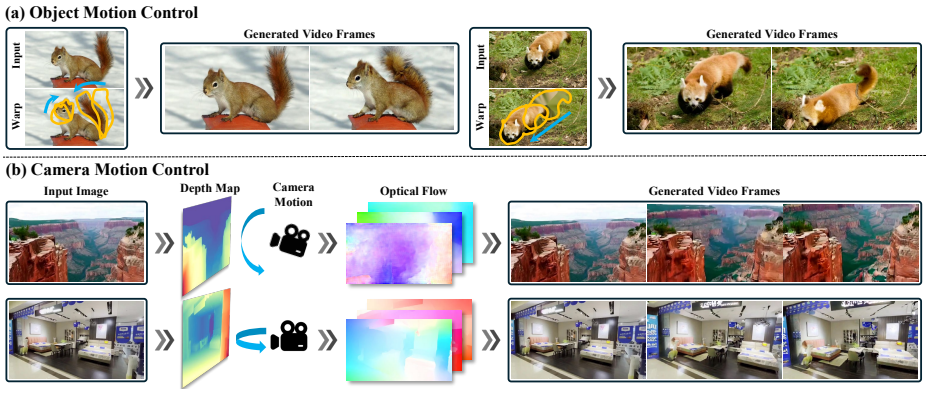

Fig. 1: We present QWERTY, a *training-free* framework that controls the motion of an image-to-video diffusion transformer (DiT) according to user-defined warping. By warping the queries of the DiT at inference time, QWERTY steers motion while preserving the generative fidelity of the pretrained backbone. Our framework enables (a) fine-grained object control by translating, rotating, and scaling masks drawn on the input image, and (b) supports camera motion control by conditioning on optical flow.

Abstract. Video diffusion transformers (DiTs) generate high-fidelity and temporally coherent videos, yet motion control remains implicit, primarily relying on text prompts. As a result, achieving desired motion often requires extensive prompt engineering and repeated resampling. While fine-tuning models with additional spatial prompts (e.g., bounding boxes or point trajectories) enables explicit control, it demands substantial data curation and computation, and may compromise the generative capabilities of pretrained models. Consequently, training-free motion control using such spatial prompts has been explored in U-Net–based video diffusion models, but remains largely unexplored for DiTs. We introduce QWERTY, a training-free framework that enables flexible motion control in pretrained image-to-video DiTs via user-defined object warping and optical flow. We carefully manipulate the 3D full attention of DiTs

⋆ Corresponding author

by warping the frame-invariant semantic subspace of queries. We find that the noise predicted by the query-warped DiT naturally guides the diffusion trajectory toward the desired motion, and further show that leveraging this noise as self-guidance for latent optimization improves control stability and visual quality. Experiments show that QWERTY achieves the most effective motion control among existing training-free approaches on a recent image-to-video DiT, with performance comparable to fine-tuning–based methods.

Keywords: Diffusion transformer · Training-free · Video motion control

1 Introduction

Recent advances in video diffusion transformers (DiTs) enable the generation of high-quality videos with longer, more dynamic, and natural motion [9, 28]. However, current pretrained models still control motion primarily through text prompts (*e.g.*, *make the squirrel wag its tail*), which are fundamentally ill-suited to precisely specifying where, how far, and how fast things should move [7, 26]. This mismatch has driven growing demand for more explicit motion control.

In response, many methods specify motion via spatial prompts such as masks [6, 37], bounding boxes [17, 26, 48], and point trajectories [58], and more recent approaches introduce richer representations like landmarks [27], free-form warping fields [4] and multiple fine-grained trajectories [7]. To inject such structural cues into pretrained models, existing methods typically add ControlNet-style [55] branches [7], warp the input noise [4], or encode the cues as extra tokens [57] and finetune the model. However, these approaches require resource-intensive training for each new model and can even degrade the generative capability of the pretrained backbone, which substantially undermines their practical usability.

In light of these limitations, prior work on U-Net–based video diffusion models has explored training-free motion control [12, 13, 18, 46, 47, 51]. Promising strategies developed for U-Nets—such as noise warping [32], attention masking [15], and latent optimization based on bounding box similarity [26, 48]—are either incompatible with the 3D full attention of DiTs or fail to function reliably in practice (Sec. 5). More recent training-free motion transfer methods built on DiTs mimic motion from a reference video [8, 31, 36]. However, many practical scenarios require users to specify motion directly rather than relying on reference videos, highlighting the need for flexible, user-defined motion control in DiTs.

In this work, we introduce QWERTY, a training-free framework that enables user-defined control of object and camera motion in image-to-video (I2V) DiTs using mask warping or optical flow. Consistent with prior findings that temporal correspondences in DiTs are more salient in attention maps than in hidden states [25,31], our method steers motion generation by directly manipulating attention. As shown in the second row of Fig. 2(a), the first-frame features of an I2V DiT exhibit a clear spatial layout even at early denoising steps, making them a reliable *reference* feature map. We therefore warp the regions of first-frame queries corresponding to the desired motion and paste them into later frames,

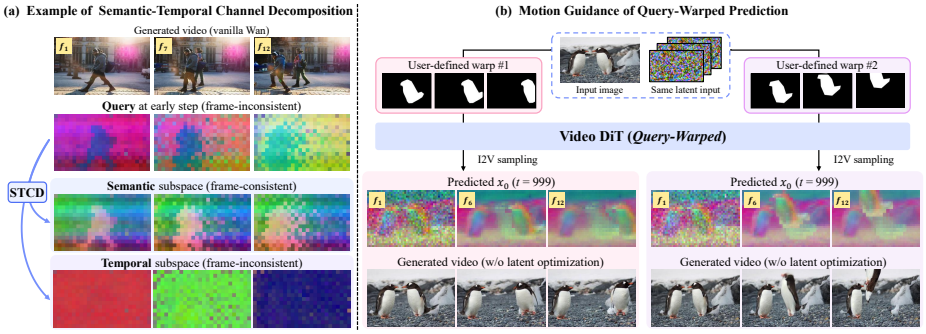

Fig. 2: Illustration of the two key components in our *training-free* video motion control pipeline. (a) Our channel decomposition technique disentangles frame-inconsistent DiT query features into semantic and temporal subspace. The filtered semantic channels are frame-consistent and spatially discriminative, making them amenable to explicit motion control. (b) We warp queries to manipulate the 3D full attention of a video DiT, thereby steering motion generation. Even at an early timestep (t=999), starting from the same input latent, the query-warped DiT already provides distinct motion guidance that follows each user-specified warp, without any additional optimization.

encouraging high attention between query tokens at frame n and their matched key tokens from the first frame. Through this query-warping mechanism, we induce the desired correspondences within the 3D full attention, and demonstrate theoretically (Sec. 3) and empirically (Sec. 5.3) that warping queries is the only effective mechanism for achieving this behavior among all attention components.

However, consistent with observations in prior work [26, 36], we find that DiT features are not frame-consistent (see the frame-variant query colors in Fig. 2(a)). These frame-inconsistent components emerge after applying temporal rotary positional embeddings (RoPE) [39], and appear to encode the temporal ordering of frames rather than scene-level semantics. Consequently, pasting such queries across frames can distort the attention distribution and introduce visual artifacts. To address this issue, we propose semantic–temporal channel decomposition (STCD). STCD computes channel-wise saliency scores by performing principal component analysis (PCA) separately on semantic and temporal anchor tokens, and then partitions the query channels into two groups based on these scores (Sec. 4.1). As shown in Fig. 2(a), this decomposition yields a frame-consistent semantic subspace and a frame-variant temporal subspace within the query. We apply query warping only to the former, enabling controlled motion generation while minimizing distortion of the attention distribution (Sec. 4.2).

We verify that the noise predicted by a query-warped DiT steers the diffusion path toward the intended motion. As shown in Fig. 2(b), at the first denoising step ($t = 999$), the predicted $\mathbf{x}_0$ derived from the query-warped noise already exhibits the desired layout according to the user-defined warp. Starting from the same latent, different warp configurations yield videos with motions consistent with the applied warps, indicating that query warping provides strong motion

guidance in the early denoising steps, where global motion patterns form [52]. While this query-warped denoising alone can guide motion, we further improve control stability and visual quality by optimizing the input latent using query-warped noise as guidance (Sec. 4.3). By integrating these components, we reveal an implicit motion prior encoded in the 3D full attention maps of DiTs, which may serve as a key ingredient for practical training-free motion control.

Contributions. Our main contributions are as follows:

- We present QWERTY, the first training-free motion control framework specifically designed for image-to-video DiTs, supporting flexible user-defined motion control via mask warping for objects and optical flow for cameras.
- We show that our query warping enables *explicit* manipulation of the 3D full attention in DiTs, and that the query-warped noise not only steers the diffusion path toward desired motion, but also serves as guidance for latent optimization, establishing a new mechanism for motion control in DiTs.
- We propose semantic–temporal channel decomposition (STCD), which enables motion control via query warping with minimal distortion of the attention distribution, thereby preserving visual quality. The anchoring mechanism is generally applicable for refining frame-inconsistent video features.
- Extensive experiments show that QWERTY achieves state-of-the-art performance among training-free methods for object and camera motion control on DiTs, substantially narrowing the gap to fine-tuning–based approaches.

2 Related Works

Video Diffusion Models. Following the success of text-to-image diffusion [33–35], early studies extended diffusion to video generation using U-Net-based architectures [3, 5, 49]. These methods typically combine spatial attention within frames and temporal attention across tokens at the same spatial location to model video dynamics. However, this factorized design prevents direct interaction between tokens at different spatial locations across frames within a single operation, limiting the modeling of globally coherent motion. More recently, diffusion transformer (DiT) [30] based video models employ 3D full attention to jointly model all tokens across space and time, enabling stronger temporal coherence [16, 23, 42, 50]. Motivated by this architectural shift toward DiT backbones, our query-warping mechanism leverages the 3D full-attention structure of DiTs.

Motion-controllable Video Generation. To enhance the practical usability of video generation, it is important to directly control desired motion rather than merely producing plausible dynamics. Early approaches relied on intuitive spatial prompts such as bounding boxes [13, 15, 54] or 2D trajectories [38, 46, 53, 56, 57]. Later works introduced warping-based control [4, 44] and sparse point trajectory prompting [7, 17], enabling more expressive motion specification. In addition, 3D motion trajectories [10, 44, 45, 47] have been explored to provide depth-aware controllability, and recent work further extends this direction by

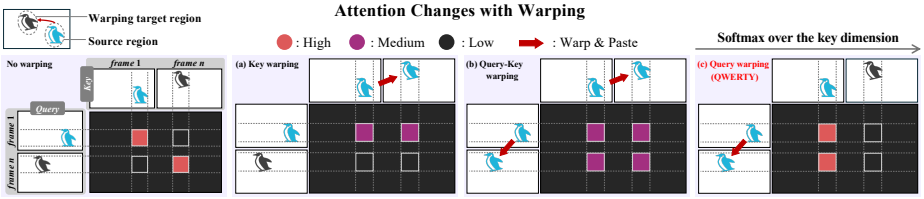

Fig. 3: A conceptual illustration of how query warping guides motion. Warping takes tokens from the object's source region in frame 1 and pastes them at warped positions in frame n. Because the softmax is applied over keys, warping only the keys (a) or both queries and keys (b) does not establish a clear correspondence between the source and target regions. In (c), query warping concentrates attention from the target region in frame n onto the source region in frame 1, which carries reliable semantic features.

supporting heterogeneous motion inputs [18]. Motion transfer [8, 21, 29, 31, 36] offers another practical route for controllable generation, but relies on reference videos and therefore does not allow users to directly specify motion.

Training-free Video Motion Control. While motion control can be achieved by fine-tuning pretrained video models, the high retraining cost has driven growing interest in training-free approaches. Methods supporting user-defined motion typically induce motion by modifying the input latent, either through latent warping [15, 32] or by optimizing the latent with bounding-box similarity losses on intermediate features [26, 48, 51]. Another line of work focuses on motion transfer, matching motion representations from attention blocks to those of a reference video [21, 31, 52]. Recent motion-transfer approaches further manipulate attention components by injecting reference queries [1,2] or modifying rotary position embeddings (RoPE) [8]. Unlike these methods, which inherit attention correspondences from a reference video, we explicitly warp queries to construct such correspondences for the desired motion. To the best of our knowledge, few prior works directly manipulate DiT attention to generate videos from a single image while following user-defined object or camera motion.

3 Why Query Warping?

This section analyzes why warping only the queries, among the attention components (queries, keys, and hidden states), is the most effective choice. Based on the observation in Fig. 2(a), we consider the first frame to be the most stable feature map during early denoising steps, and therefore a reliable reference. Thus, the goal is to increase the attention between the target region in frame n and the source region in frame 1. This encourages the target region to receive value vectors from the first-frame source region, thereby guiding the target motion. Fig. 3 illustrates how attention changes under three different warping strategies:

(a) Key warping. When keys are warped, the target-region query in frame n fails to attend strongly to any key, resulting in a dispersed attention pattern.

Moreover, the attention of the source-region query in frame 1 is also dispersed, yielding no beneficial effect. Since softmax is applied over the key dimension independently for each query, modifying keys is not suitable for concentrating the attention of multiple queries onto a single meaningful source region.

(b) Query–Key warping (hidden-state warping). Since queries and keys are derived from the same hidden state, jointly warping them behaves similarly to warping the hidden state itself. Although the attention between the target-region query in frame n and the source-region key in frame 1 increases, attention is still spread out, similar to the key-warping case, and guidance remains diluted.

(c) Query warping (QWERTY). Warping the source-region queries from the first frame to the target region in frame n yields strong attention between those queries and the first-frame source-region keys. Thus, warping only the queries pulls the target-region features toward the reference features, providing clear motion guidance. Our results in Sec. 5.3 empirically support this design choice.

This intuition can be formalized via the attention logits under a mild self-correspondence assumption. Motion control requires a target token i in frame n to attend to a source token s in frame 1. The target-to-source attention is $\alpha_{i,s} = \exp(q_i^\top k_s/\sqrt{d})/\sum_j \exp(q_i^\top k_j/\sqrt{d})$. Assuming that the source query aligns most with its own key, *i.e.*, the self-correspondence score $c_{\text{self}} \triangleq q_s^\top k_s$ satisfies $c_{\text{self}} > q_s^\top k_j,\ \forall j \neq s$, query warping sets $q_i \leftarrow q_s$, so $q_i^\top k_s \approx c_{\text{self}}$, directly increasing the numerator. In contrast, key warping only replaces target keys with source-like keys, *e.g.*, $\tilde{k}_t \leftarrow k_s$, leaving the numerator $q_i^\top k_s$ unchanged; thus, $\alpha_{i,s}$ need not increase. Query-key warping sets $q_i \leftarrow q_s$, like query-only warping, but also replaces target keys with source-like keys, $\tilde{k}_t \leftarrow k_s$. Thus, query-key warping also increases the numerator, but adds strong competing denominator terms with scores near c_{self}, diluting attention to the source key k_s, unlike query-only warping. This formalizes why query warping most directly increases target-to-source attention and induces the desired correspondence.

4 Methods

In this section, we present QWERTY, as outlined in Fig. 4. First, we introduce STCD (Sec. 4.1). Second, we detail the query-warping procedure, including the application of STCD to queries (Sec. 4.2). Finally, we describe the motion-guided generation pipeline that performs diffusion with the query-warped DiT and latent optimization (Sec. 4.3). Although we describe our method using Wan 2.2 [42], it can be applied to standard video DiTs with similar 3D full attention.

4.1 Semantic-Temporal Channel Decomposition

STCD is a PCA-driven disentanglement technique that isolates a frame-consistent semantic subspace from DiT features whose channels are entangled with temporal dynamics (Fig. 2(a), second row). We first define two anchor token sets: *semantic anchors*, whose within-frame spatial structure emphasizes semantic

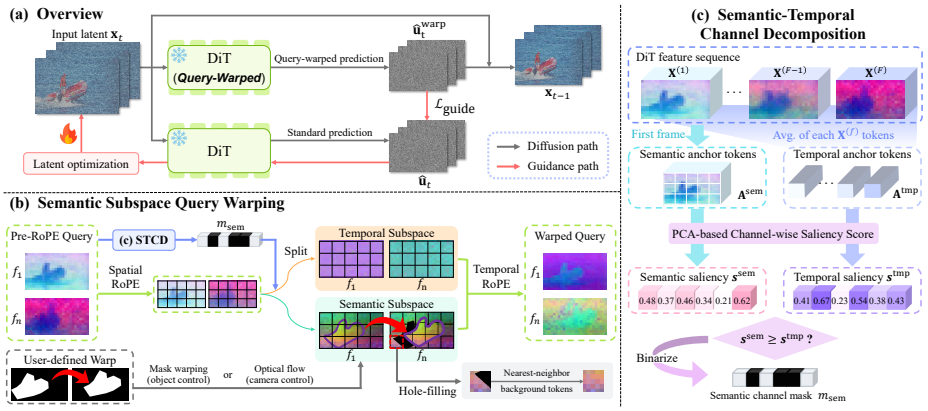

Fig. 4: Pipeline overview. (a) We perform diffusion steps using motion-inducing noise predicted from the query-warped video DiT. This query-warped noise is also used as a guidance signal to optimize the input latent, further stabilizing control. (b) We isolate frame-consistent semantic channels from the queries and warp them together with only the spatial axis of the 3D RoPE, delicately reshaping the attention distribution to induce motion without excessive distortion. (c) To separate semantic and temporal components in the query features, we define semantic anchor tokens and temporal anchor tokens from the query tokens, perform PCA on each anchor set to obtain channel-wise saliency scores, and then split the channels into two subspaces based on these scores.

content, and *temporal anchors*, whose across-frame variability highlights temporal change. Running PCA on each anchor set yields channel-wise saliency scores. We then compare the scores for each channel and apply query warping only where semantic saliency exceeds temporal saliency (Sec. 4.2). Operationally, STCD defines a function that, given a feature tensor, returns a mask selecting semantically salient channels. The overall procedure is illustrated in Fig. 4(c).

PCA-based Channel Saliency. We begin by explaining how, given an arbitrary anchor tokens $\mathbf{A} \in \mathbb{R}^{N \times D}$ with N tokens and D channels, we compute a channel-wise *discriminativeness* score. Intuitively, a channel is discriminative for $\mathbf{A}$ if its values vary systematically across tokens (*e.g.*, across spatial locations within a frame or across frames). PCA summarizes this cross-token variation: components with higher variance are more informative for distinguishing tokens.

Concretely, we center $\mathbf{A}$ across tokens and take its singular value decomposition (SVD), $\overline{\mathbf{A}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where the columns of $\mathbf{V}$ are channel loadings and the diagonal entries of $\mathbf{\Sigma}$ are the corresponding singular values σ_i. We then assign each channel d a saliency by weighting its loading magnitude by the corresponding singular values over the top k components:

$$s_d(\mathbf{A}, k) = \max_{1 \leq i \leq k} \sigma_i \left| V_{d,i} \right|, \qquad d = 1, \ldots, D. \tag{1}$$

This yields a score vector $\boldsymbol{s}(\mathbf{A}, k) \in \mathbb{R}^D$ that identifies, for the given anchor token set, the channels most responsible for dominant token-wise variation. We

compute $\boldsymbol{s}$ for both the semantic and temporal anchor sets and compare them per channel to label channels as semantic or frame-variant.

Semantic Channel Saliency. We treat the first frame's token features, $\mathbf{X}^{(1)} \in \mathbb{R}^{P\times D}$, where P denotes the number of tokens in a single frame, as the *semantic anchor* set, *i.e.*, $\mathbf{A}^{\text{sem}} = \mathbf{X}^{(1)}$. In DiT-based I2V models, first-frame features already exhibit a clear, well-structured layout even at early denoising steps (Fig. 2(a)). This variation across spatial locations provides strong semantic cues, making $\mathbf{X}^{(1)}$ a suitable semantic anchor. We apply the saliency function in Eq. (1) to obtain the semantic channel score vector

$$\boldsymbol{s}^{\text{sem}} = \boldsymbol{s}\big(\mathbf{A}^{\text{sem}},\, k_{\text{sem}}\big) \in \mathbb{R}^{D}. \tag{2}$$

Here, k_{sem} denotes the number of principal components.

Temporal Channel Saliency. To capture channels that vary across frames, we form *temporal anchor* tokens by averaging over spatial positions within each frame. This suppresses within-frame spatial idiosyncrasies while preserving across-frame trends. Concretely, we compute $\mathbf{a}_f^{\text{tmp}} = \frac{1}{P}\sum_{p=1}^{P} \mathbf{X}_{p,:}^{(f)} \in \mathbb{R}^{D}$ for each frame f, and stack them into $\mathbf{A}^{\text{tmp}} \in \mathbb{R}^{F\times D}$ with rows $\mathbf{A}_{f,:}^{\text{tmp}} = \mathbf{a}_f^{\text{tmp}}$. Applying the same scoring operator yields the temporal channel saliency vector

$$\boldsymbol{s}^{\text{tmp}} = \boldsymbol{s}\big(\mathbf{A}^{\text{tmp}},\, k_{\text{tmp}}\big) \in \mathbb{R}^{D}. \tag{3}$$

Frame-consistent Semantic Channel Masking. We identify channels that are more semantically salient than temporally salient by comparing the two saliency scores channel-wise (Eqs. (2) and (3)). Specifically, we form a semantic channel mask $m_{\text{sem},d} = \mathbb{1}[s_d^{\text{sem}} \geq s_d^{\text{tmp}}]$. To restrict query warping to semantic channels, we define a function that maps a feature sequence $\mathbf{X}$ to the binary semantic channel mask:

$$\boldsymbol{m}_{\text{sem}} \;=\; \text{STCD}\big(\mathbf{X};\, k_{\text{sem}}, k_{\text{tmp}}\big) \in \{0,1\}^{D}. \tag{4}$$

4.2 Semantic-Subspace Query Warping

We first briefly review 3D RoPE for clarity, and then describe the query warping procedure, along with our strategy for handling holes that arise during warping.

Preliminary: 3D RoPE of video DiTs. Rotary positional embedding (RoPE) [39] injects relative positional information into self-attention by rotating query and key feature dimensions. After rotation, the attention dot product depends on relative displacement. For video, 3D RoPE treats each token as (x, y, t). It splits the channel dimension into $[D_x \,\|\, D_y \,\|\, D_t]$, applies a separate 1D RoPE to each block, and then concatenates the rotated blocks. Because these 1D RoPEs are applied to disjoint channel blocks, we can apply the spatial and temporal RoPE separately in our query warping.

Query Warping Pipeline. As illustrated in Fig. 4(b), our query-warping pipeline proceeds in three steps. **(i)** From the pre-RoPE queries, we compute the frame-consistent semantic channel mask via STCD (Eq. (4)). **(ii)** We apply the spatial RoPE to the queries and warp the first-frame queries onto later frames according to the desired motion, pasting only the channels selected by the STCD mask. By performing warping after spatial RoPE, the queries in frame n within the warped mask region share the same spatial positional encoding as the corresponding first-frame keys, thereby strengthening their spatial affinity in attention. Applying temporal RoPE jointly during warping would further increase their attention scores; however, this would encode the mask region in frame n with the temporal position of frame 1, leading to excessive distortion of the attention distribution. **(iii)** Finally, we apply the temporal RoPE to the warped queries and proceed with the standard attention and denoising steps. Through this carefully designed query-warping strategy, we guide motion without disrupting the internal dynamics of the pretrained model.

Handling Warping Holes. We detail the hole-handling strategies during warping, which improve control fidelity and suppress visual artifacts. Additional analysis is provided in the Supplementary Material.
(i) Warping with a mask (object control). Simply pasting the warped mask region leaves residual object traces at the source location; we therefore explicitly handle the source-region hole. On the first frame, we fill the hole with nearest-neighbor background tokens and paste both the warped mask region and the filled-hole region into each target frame (Fig. 4(b)). This suppresses cloning artifacts that would otherwise leave a duplicate at the source location.
(ii) Warping with optical flow (camera control). Since optical flow is defined over the full spatial domain of each frame, warping is applied to the entire frame. Unlike the mask case, flow-based forward warping inherently creates holes on the target sampling grid due to incomplete coverage. Camera motion can also reveal regions not visible in the first frame, producing additional holes. Filling these holes with background tokens (akin to backward warping) propagates neighboring values into empty regions, causing stretching artifacts. Instead, we do not paste warped queries at hole locations; we leave the corresponding target-frame queries unchanged and allow the model to inpaint them naturally.

4.3 Motion-guided Generation via Query Warping

In this section, we describe the overall framework that integrates the components from Secs. 4.1 and 4.2 to control video motion through query-warped diffusion sampling and latent optimization, as illustrated in Fig. 4(a).

Query-Warped Prediction. We describe how query warping guides the I2V DiT under flow-matching formulation [22] and note that the same principle applies to standard diffusion-based sampling. Let $\mathbf{c}$ denote the input image condition and $\mathbf{x}_t$ the latent at timestep t. The DiT, parameterized by θ, predicts the

velocity field. We denote the *standard* velocity prediction by

$$\hat{\mathbf{u}}_t = f_\theta(\mathbf{x}_t, t, \mathbf{c}). \tag{5}$$

The latent state is then advanced by an Euler step of the underlying flow:

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t + \Delta t\, \hat{\mathbf{u}}_t. \tag{6}$$

To incorporate user-defined motion, we apply the query-warping procedure of Sec. 4.2 within the DiT attention blocks. Let $f_\theta^{\text{warp}}(\mathbf{x}_t, t, \mathbf{c}; \mathcal{W})$ denote a forward pass in which queries are warped according to the mask or flow field $\mathcal{W}$. This yields the *query-warped* prediction:

$$\hat{\mathbf{u}}_t^{\text{warp}} = f_\theta^{\text{warp}}(\mathbf{x}_t, t, \mathbf{c}; \mathcal{W}). \tag{7}$$

At early timesteps (i.e., the first five denoising steps), we use $\hat{\mathbf{u}}_t^{\text{warp}}$ in place of $\hat{\mathbf{u}}_t$ in Eq. (6), biasing the trajectory toward the desired motion.

Self-guided Latent Optimization. While sampling with the query-warped prediction already provides strong motion guidance (Fig. 2(b)), we further leverage it as a self-guidance signal for latent optimization. By directly updating the input latent, we can more explicitly steer the diffusion trajectory toward the desired motion. Specifically, at each guided timestep, we compute both the standard prediction $\hat{\mathbf{u}}_t$ and the query-warped prediction $\hat{\mathbf{u}}_t^{\text{warp}}$. Keeping the DiT parameters frozen, we update the latent $\mathbf{x}_t$ to match the standard prediction to the warped one. Gradients are stopped on the warped branch and propagated only through the standard branch (Fig. 4(a)).

To stabilize the guidance, we adopt the phase-constraint idea of RoPECraft [8], encouraging the standard and query-warped predictions to match in magnitude while aligning their Fourier phases, which capture motion structure. We formulate the self-guided loss as

$$\begin{aligned}\mathcal{L}_{\text{guide}} = \|\hat{\mathbf{u}}_t - \hat{\mathbf{u}}_t^{\text{warp}}\|_2^2 + \lambda \,\|\cos(\angle\mathcal{F}(\hat{\mathbf{u}}_t)) - \cos(\angle\mathcal{F}(\hat{\mathbf{u}}_t^{\text{warp}}))\| \\ + \lambda \,\|\sin(\angle\mathcal{F}(\hat{\mathbf{u}}_t)) - \sin(\angle\mathcal{F}(\hat{\mathbf{u}}_t^{\text{warp}}))\|\end{aligned} \tag{8}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform, $\angle$ extracts the phase angle, and λ controls the weight of the phase constraint. Following RoPECraft, we set $\lambda = 0.1$.

5 Experiments

In this section, we present experiments to validate Qwerty. Additional results, analyses, and discussions are provided in the supplementary material.

5.1 Setup

Baselines. We compare Qwerty with both U-Net– and DiT-based motion-control methods. SG-I2V [26] and MOFT [48] are training-free U-Net approaches

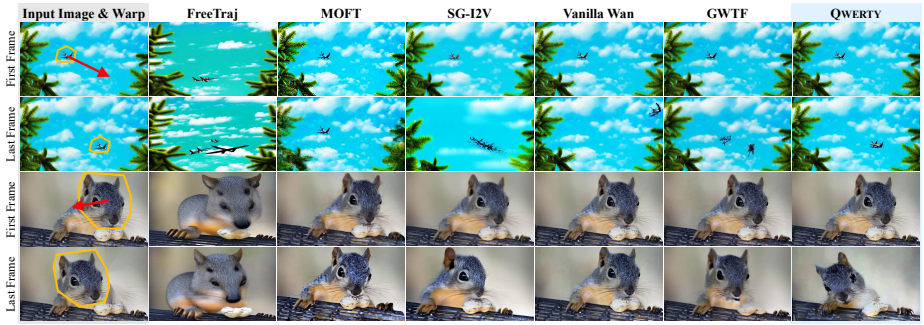

Fig. 5: Qualitative comparisons of local *object* motion control. Polygonal object region masks and user-defined warps are given as input.

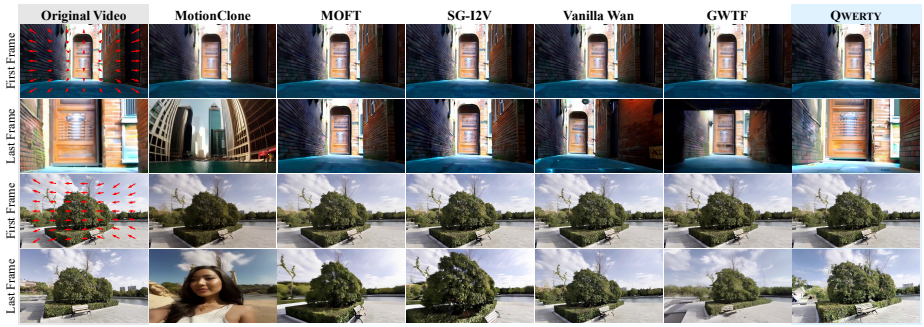

Fig. 6: Qualitative comparisons of *camera* motion control. For evaluation, we use optical flow estimated from the original video. In practice, users can instead provide optical flow derived from a depth map and camera trajectory (Sec. 5.3).

that use bounding-box trajectories to control object or camera motion. FreeTraj [32], originally designed for text-to-video (T2V), is adapted to the I2V setting and evaluated only for object control since it does not support camera motion. MotionClone [21], a T2V motion-transfer method, is also adapted to I2V, where we provide the ground-truth video as motion reference to establish a strong U-Net baseline for camera control comparison. Since direct training-free counterparts for DiT-based I2V models are not readily available, we adapt methods developed for U-Net models, including Noise Warping [32] and SG-I2V, to the DiT backbone. We include Go-With-The-Flow (GWTF) [4], a finetuned DiT model enabling motion control with the same mask and optical-flow conditioning interface as ours, serving as a strong upper-bound baseline. Our main results are reported on Wan 2.2 TI2V-5B [42], with additional ablations on CogVideoX-I2V-5B to demonstrate cross-backbone effectiveness. Implementation details for all baselines and our method are provided in the supplementary material.

Datasets. To evaluate both video quality and motion controllability, we conduct experiments on two datasets targeting object and camera motion control. Fol-

Table 1: Quantitative comparisons with baselines. ⇑/⇓ indicates better is higher/lower. **Best** and second best are marked. Light blue rows denote our model. * indicates using the ground-truth video as a motion reference for motion transfer.

	Method	Training free	Backbone	FID ⇓	FVD ⇓ ($\times 10^3$)	FTD ⇓	VBench ⇑			
							SC	BC	MS	TF
	Object motion control (VIPSeg)									
U-Net	FreeTraj	✓	VideoCrafter	126.8	0.924	0.528	0.886	0.901	0.979	0.968
U-Net	MOFT	✓	SVD	76.04	0.861	0.527	0.955	0.944	0.981	0.964
U-Net	SG-I2V	✓	SVD	60.83	0.798	0.579	0.952	0.961	0.974	0.946
DiT	Vanilla Wan	–	Wan	46.17	0.677	0.534	0.977	0.973	0.983	0.973
DiT	Noise Warping	✓	Wan	270.38	3.392	0.569	0.725	0.753	0.955	0.955
DiT	SG-I2V	✓	Wan	107.98	0.704	0.600	0.976	0.960	0.962	0.967
DiT	GWTF	✗	–	**40.19**	**0.645**	0.528	0.958	0.975	**0.993**	**0.988**
DiT	Qwerty	✓	Wan	45.95	0.667	**0.512**	**0.978**	**0.981**	0.985	0.977
	Camera motion control (DL3DV)									
U-Net	MotionClone*	✓	AnimateDiff	96.63	1.624	0.683	0.897	0.934	0.969	0.931
U-Net	MOFT	✓	SVD	51.84	1.327	0.424	0.955	0.967	0.958	0.961
U-Net	SG-I2V	✓	SVD	45.18	1.265	0.578	0.966	0.973	0.979	0.936
DiT	Vanilla Wan	–	Wan	32.21	1.362	0.367	**0.977**	0.974	0.978	0.967
DiT	Noise Warping	✓	Wan	282.93	5.663	0.455	0.719	0.747	0.972	0.966
DiT	SG-I2V	✓	Wan	42.46	1.524	0.438	0.976	0.970	0.981	0.941
DiT	GWTF	✗	–	**30.32**	1.576	0.513	0.962	**0.978**	**0.988**	**0.976**
DiT	Qwerty	✓	Wan	35.56	**1.007**	**0.237**	0.946	0.956	0.941	0.935

lowing prior works [4, 26], object motion control is evaluated on 100 VIPSeg [24] validation videos with dominant object motion, where mask-based warps are annotated. Camera motion control is evaluated on 100 randomly sampled videos from DL3DV [20]. Preprocessing and labeling details are provided in the supplementary material, and the selected video IDs will be released with the code.

Evaluation metrics. We evaluate both generation quality and motion controllability of the generated videos. For generation quality, we report Fréchet Inception Distance (FID) [11], Fréchet Video Distance (FVD) [41], and four VBench metrics [14, 59]: subject consistency (SC), background consistency (BC), motion smoothness (MS), and temporal flickering (TF). For motion controllability, we adopt the recent Fréchet Trajectory Distance (FTD) [8], which measures alignment with the target motion for both object and camera movement. For fair comparison, videos are generated at their native resolution and resized to 480×832 for evaluation, and the frame count is standardized by truncating each video to the smallest maximum length supported by the compared models.

5.2 Qualitative and Quantitative Results

We present qualitative and quantitative comparisons with baselines in Fig. 5, Fig. 6, and Table 1. Fig. 7 shows training-free control methods designed for U-Net do not transfer well to DiT. Our results are mainly reported on Wan, while qualitative results on CogVideoX are provided in the supplementary material.

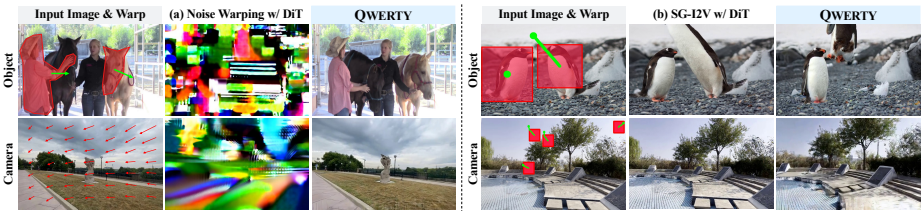

Fig. 7: Failure of U-Net–based motion-control methods on DiT. (a) Noise Warping collapses generation. (b) SG-I2V yields mostly static results, with occasional unintended motion. In contrast, QWERTY achieves reliable object and camera control on DiT.

Object control. Fig. 5 illustrates strong object motion controllability of our method. Baseline approaches often exhibit limited motion quality, producing duplicated or distorted objects. Some methods fail to preserve consistency with the input image, while others incorrectly induce camera movement instead of object motion. Although GWTF, a training-based approach, tends to follow the control more strictly, the resulting motion is often contextually unnatural or distorted. In contrast, our method generates high-fidelity videos with accurate motion compliance while producing semantically coherent motion. This behavior is more clearly observed in the supplementary videos. Consistently, Table 1 shows that our method achieves the best overall performance among training-free approaches and delivers video quality and control accuracy comparable to GWTF. Notably, despite manipulating internal components of the model for motion control, our approach improves visual fidelity compared to vanilla Wan.

Camera control. Fig. 6 shows that QWERTY aligns most faithfully with the instructed motion prompts, even under challenging conditions such as large zoom-ins or tilting camera motions. In contrast, other methods either fail to follow the specified motion or produce noticeable artifacts. In particular, GWTF performs poorly on DL3DV, a real-world scene dataset, suggesting that fine-tuning–based approaches may degrade the backbone's pretrained generalization. In Table 1, QWERTY outperforms nearly all baselines in camera-motion controllability, achieving substantial improvements in both FVD and FTD. Since generating camera motion inherently requires synthesizing novel scene content, VBench metrics such as SC or BC tend to trade off with camera-motion controllability. Methods such as vanilla Wan or SG-I2V adapted to DiT achieve higher VBench scores, as they often produce static or uncontrolled videos rather than true motion control. Notably, our method surpasses the training-based GWTF in both FVD and FTD, highlighting the effectiveness of our training-free design.

U-Net control methods on DiT. Fig. 7 shows training-free motion-control strategies originally designed for U-Net, such as Noise Warping and SG-I2V, do not transfer well to DiT. Noise Warping produces visually collapsed videos for both object and camera motion control (Fig. 7(a)), which is reflected in its poor FID, FVD, and VBench scores in Table 1. SG-I2V mostly generates static videos for both object and camera control, with only occasional responses to the control

Table 2: Ablation results on Wan for camera control on the DL3DV dataset. ⇑/⇓ indicates better is higher/lower. **Best** and second best are highlighted.

	Option	FVD ⇓ ($\times 10^3$)	FTD ⇓	VBench ⇑ BC	VBench ⇑ MS
	Vanilla Wan	1.362	0.367	**0.974**	0.978
Warping Target	(a) No STCD	1.348	0.297	0.956	0.940
	(b) *w/o* spat. RoPE	1.419	0.299	0.961	0.978
	(c) *w/* temp. RoPE	1.348	0.285	0.955	0.945
	(d) K	1.482	0.293	0.969	0.861
	(e) Q, K	1.427	0.290	0.971	**0.980**
	(f) Attn. output	1.590	0.288	0.940	0.939
Optim.	(g) No opt.	1.301	0.274	0.962	0.956
	(h) Opt. only	1.423	0.355	0.940	0.939
	(i) *w/o* phase loss	1.440	0.285	0.949	0.942
	QWERTY	**1.007**	**0.237**	0.956	0.941

Table 3: Ablation results on CogVideoX for object control on the VIPSeg dataset. ⇑/⇓ indicates better is higher/lower. **Best** and second best are highlighted.

	Option	FVD ⇓ ($\times 10^3$)	FTD ⇓	VBench ⇑ BC	VBench ⇑ MS
	Vanilla CogVideoX	0.669	0.537	0.974	0.990
Warping Target	(a) No STCD	0.697	0.542	0.961	0.987
	(b) *w/o* spat. RoPE	1.578	0.531	0.835	0.973
	(c) *w/* temp. RoPE	0.831	0.511	0.959	0.944
	(d) K	0.759	0.544	0.966	0.928
	(e) Q, K	0.727	0.531	0.975	0.989
	(f) Attn. output	0.745	0.544	0.943	0.971
Optim.	(g) No opt.	0.669	0.501	0.972	0.980
	(h) Opt. only	0.688	0.536	0.977	0.979
	(i) *w/o* phase loss	0.704	0.514	0.964	0.955
	QWERTY	**0.668**	**0.473**	**0.980**	**0.993**

signals. As illustrated in Fig. 7(b), the penguin's torso is abnormally stretched along the bounding-box trajectory, suggesting that partial guidance effective in U-Net fails on DiT due to the weaker locality bias in DiT. Consequently, SG-I2V exhibits low FID and FVD in Table 1, while producing relatively high VBench scores since the generated videos are largely static with little frame-wise variation. Moreover, methods like SG-I2V require ad-hoc search over guidance blocks, which becomes impractical for DiT models with many blocks. In contrast, our method applies query warping across all blocks with negligible overhead and injects guidance only at the final output, eliminating the need for block selection. Further analysis of SG-I2V on DiT is provided in the supplementary material.

5.3 Ablation Studies

We conduct extensive ablations to validate each design choice in our framework. Table 2 and Table 3 present ablation results on Wan with DL3DV and CogVideoX with VIPSeg, respectively. Together, results demonstrate that our design is effective across both backbones and motion types. Additional ablation videos and hyperparameter studies are provided in the supplementary material.

Warping target. We first ablate the components involved in query warping. Removing STCD (a), omitting spatial RoPE warping (b), or additionally warping temporal RoPE (c) all degrade performance, confirming that selectively warping only the semantic component is essential for stable control. We also examine warping other attention components. Warping the keys only (d), warping both queries and keys (e), or warping the attention output (f) all underperform compared to our query-only strategy. The relatively high VBench scores for (e) and Vanilla Wan stem from their tendency to produce largely static videos. This supports our theoretical analysis in Sec. 3, which suggests that query warping is uniquely suited for concentrating attention on the desired motion region.

Latent optimization. We ablate the components contributing to motion guidance. Using only query-warped denoising without latent optimization (g) yields slightly worse scores yet still retains strong control effects, indicating that query-warped denoising alone provides effective motion guidance (Fig. 2(b)). In contrast, performing latent optimization without query-warped denoising (h) results in weaker control. The performance gap from removing latent optimization is more pronounced for camera control than object control, suggesting that latent optimization can be partially omitted for object control under limited computational budgets. Finally, removing the phase-consistency constraint (i) degrades performance, highlighting its importance for motion fidelity and stability.

Effectiveness in practical scenarios. To evaluate the practicality of our camera control, we examine whether it remains effective without ground-truth motion signals. In our experiments, optical flow extracted with RAFT [40] from the ground-truth video is used as the control signal. However, in real deployment only a single input image is available, requiring optical flow to be derived from predicted depth and a user-specified camera trajectory. To simulate this setting, we estimate depth using VGGT [43] and Depth Anything 3 (DAM 3) [19], and compute depth-based optical flow using camera poses predicted by VGGT. Table 4 shows that depth-estimated flow from the two models consistently yields results comparable to those obtained with ground-truth flow. This suggests that our optical-flow-based camera control is practical for real-world use.

Table 4: Effect of optical flow source on camera control. OF err.: difference between the input and re-estimated flow from the generated video.

Option	OF err. ⇓	FTD ⇓	BC ⇑	MS ⇑
Estimated (VGGT)	0.710	0.251	0.931	0.938
Estimated (DAM 3)	0.707	0.243	0.933	0.940
Ground-truth (RAFT)	0.700	0.237	0.946	0.941

6 Conclusion

In this work, we present QWERTY, a training-free framework for user-defined control of both object and camera motion in image-to-video DiT models. Our key insight is that query features establish a stable spatial layout early in denoising, allowing motion to be steered by warping queries within the 3D full attention. We introduce STCD to isolate a frame-consistent semantic subspace that can be safely warped without disrupting attention. Query-warped predictions provide strong early-step motion guidance, enabling self-guided latent optimization that improves both motion stability and visual quality. Experiments show that QWERTY outperforms existing training-free methods and achieves performance comparable to learning-based approaches while preserving the generalization ability of the pretrained backbone. These results establish query warping as a principled mechanism for controllable DiT-based video generation.

Acknowledgement. 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.

References

1. 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)
2. Bai, J., He, T., Wang, Y., Guo, J., Hu, H., Liu, Z., Bian, J.: Uniedit: A unified tuning-free framework for video motion and appearance editing. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10171–10180 (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. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13–23 (2025)
5. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
6. Dai, Z., Zhang, Z., Yao, Y., Qiu, B., Zhu, S., Qin, L., Wang, W.: Animateanything: Fine-grained open domain image animation with motion guidance. arXiv preprint arXiv:2311.12886 (2023)
7. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Aytar, Y., Rubinstein, M., Sun, C., et al.: Motion prompting: Controlling video generation with motion trajectories. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1–12 (2025)
8. 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)
9. Google DeepMind: Veo 3: A generative video model by google deepmind. `https://aistudio.google.com/models/veo-3` (2024), accessed: 2025-11-13
10. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for video diffusion models. In: The Thirteenth International Conference on Learning Representations (2025)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
12. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024)
13. Huang, Y., Chen, Y., Ding, L., Zhang, X., Dai, W., Zou, J., Xiong, H., Tian, Q.: Im-zero: Instance-level motion controllable video generation in a zero-shot manner. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7265–7275 (2025)
14. 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)
15. 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)

16. 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)
17. Li, Q., Xing, Z., Wang, R., Zhang, H., Dai, Q., Wu, Z.: Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. arXiv preprint arXiv:2503.16421 (2025)
18. Li, Z., Luo, H., Shuai, X., Ding, H.: Anyi2v: Animating any conditional image with motion control. arXiv preprint arXiv:2507.02857 (2025)
19. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
20. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
21. 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. In: The Thirteenth International Conference on Learning Representations
22. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
23. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
24. Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., Yang, Y.: Large-scale video panoptic segmentation in the wild: A benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21033–21043 (2022)
25. 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)
26. Namekata, K., Bahmani, S., Wu, Z., Kant, Y., Gilitschenski, I., Lindell, D.B.: Sg-i2v: Self-guided trajectory control in image-to-video generation. arXiv preprint arXiv:2411.04989 (2024)
27. Niu, M., Cun, X., Wang, X., Zhang, Y., Shan, Y., Zheng, Y.: Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In: European Conference on Computer Vision. pp. 111–128. Springer (2024)
28. OpenAI: Sora: Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators/` (2024), accessed: 2025-11-13
29. 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)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
31. Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22911–22921 (2025)
32. Qiu, H., Chen, Z., Wang, Z., He, Y., Xia, M., Liu, Z.: Freetraj: Tuning-free trajectory control in video diffusion models. arXiv preprint arXiv:2406.16863 (2024)

33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
34. 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. pp. 10684–10695 (2022)
35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
36. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10995–11005 (2025)
37. 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)
38. Shi, Y., Xue, C., Liew, J.H., Pan, J., Yan, H., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8839–8849 (2024)
39. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
40. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
41. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
42. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
43. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
44. Wang, Z., Lan, Y., Zhou, S., Loy, C.C.: Objctrl-2.5 d: Training-free object control with camera poses. arXiv preprint arXiv:2412.07721 (2024)
45. 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)
46. 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)
47. Xiao, F., Liu, X., Wang, X., Peng, S., Xia, M., Shi, X., Yuan, Z., Wan, P., Zhang, D., Lin, D.: 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In: The Thirteenth International Conference on Learning Representations (2024)
48. Xiao, Z., Zhou, Y., Yang, S., Pan, X.: Video diffusion models are training-free motion interpreter and controller. Advances in Neural Information Processing Systems **37**, 76115–76138 (2024)

49. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: European Conference on Computer Vision. pp. 399–417. Springer (2024)
50. 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)
51. Yatim, D., Fridman, R., Bar-Tal, O., Kasten, Y., Dekel, T.: Space-time diffusion features for zero-shot text-driven motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8466–8476 (2024)
52. Yesiltepe, H., Meral, T.H.S., Dunlop, C., Yanardag, P.: Motionshop: Zero-shot motion transfer in video diffusion models with mixture of score guidance. arXiv preprint arXiv:2412.05355 (2024)
53. 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)
54. Yu, S., Fang, J.Z., Zheng, J., Sigurdsson, G., Ordonez, V., Piramuthu, R., Bansal, M.: Zero-shot controllable image-to-video animation via motion decomposition. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3332–3341 (2024)
55. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3813–3824. IEEE (2023)
56. Zhang, X., Duan, Z., Gong, D., Liu, L.: Training-free motion-guided video generation with enhanced temporal consistency using motion consistency loss. arXiv preprint arXiv:2501.07563 (2025)
57. 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)
58. Zhang, Z., Long, F., Qiu, Z., Pan, Y., Liu, W., Yao, T., Mei, T.: Motionpro: A precise motion controller for image-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27957–27967 (2025)
59. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)