RoPECraft: Training-Free Motion Transfer with Trajectory-Guided RoPE Optimization on Diffusion Transformers

*Ahmet Berke Gökmen[1,2] *Yiğit Ekin[1] *Bahri Batuhan Bilecen[1,3,4] Aysegul Dundar[1]

[1]Bilkent University [2]INSAIT, Sofia University "St. Kliment Ohridski" [3]ETH Zurich [4]Max Planck Institute

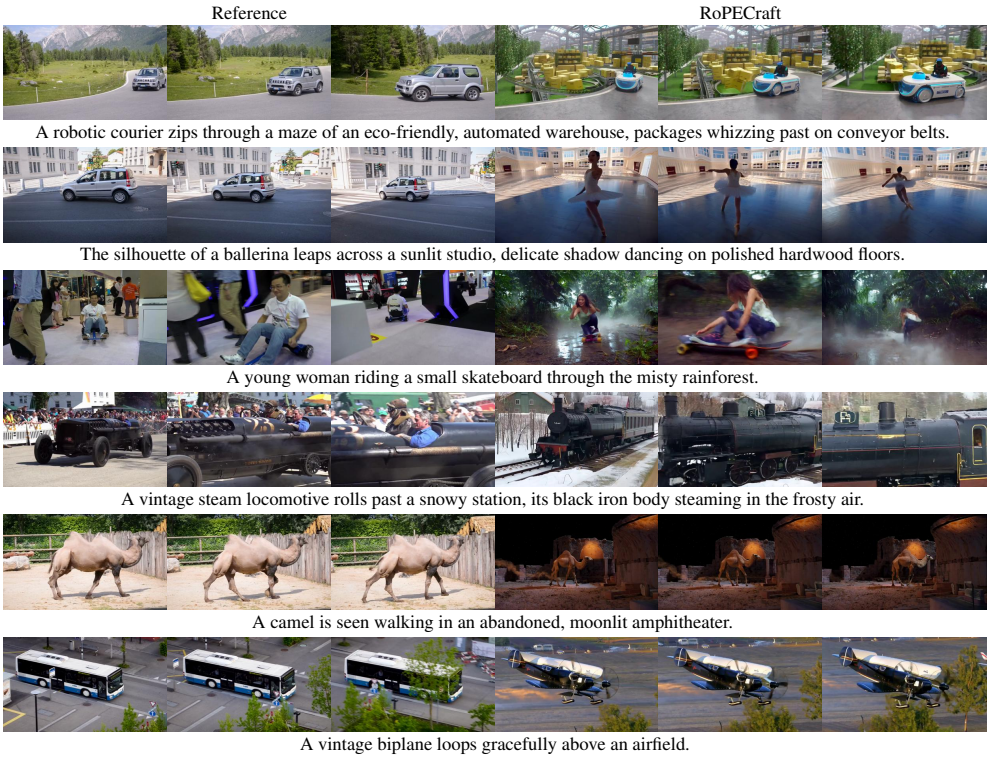


Figure 1: Our method successfully transfers the motion from reference videos.

Abstract

We propose RoPECraft, a training-free video motion transfer method for diffusion transformers that operates solely by modifying their rotary positional embeddings (RoPE). We first extract dense optical flow from a reference video, and utilize the resulting motion offsets to warp the complex-exponential tensors of RoPE, effectively encoding motion into the generation process. These embeddings are then further optimized during denoising time steps via trajectory alignment between the predicted and target velocities using a flow-matching objective. To keep the output faithful to the text prompt and prevent duplicate generations, we incorporate a regularization term based on the phase components of the reference video's Fourier transform, projecting the phase angles onto a smooth manifold to suppress high-frequency artifacts. Experiments on benchmarks reveal that RoPECraft outperforms all recently published methods, both qualitatively and quantitatively.

*Equal contribution. Correspondence: berke.gokmen@insait.ai. Project page

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Diffusion transformers (DiT) have become a leading approach for conditional video generation, producing realistic and coherent content across diverse scenarios [6, 16, 20, 50, 21, 38, 44]. While text conditioning provides a convenient interface, they are often too ambiguous to specify detailed spatio-temporal dynamics such as body movement, camera motion, or interactions. As generative quality improves, so does the demand for more precise and controllable motion synthesis.

To address the limitations of text-based motion control, earlier methods introduced explicit structural cues such as masks, bounding boxes, or depth maps to guide motion [9, 43, 40]. These approaches assume consistent geometry between the reference and generated videos, which often fails under domain shifts [45]. More recent work has shifted toward leveraging latent representations within generative models. Some methods extract motion features from internal activations [14, 45, 42], while others modify the latent prior [4] to better align reference and generated motions. A prominent example, Go with the Flow (GWTF) [4], uses a pretrained optical flow model [37] to generate motion priors that warp the initial noise input while maintaining its Gaussianity. This reportedly stabilizes and speeds up convergence. However, directly warping noise disrupts the intended latent distribution of the pre-trained DiT, as seen in Fig. 2. Even with inference-time latent optimization, the method struggles with generalization and domain shifts. As a result, GWTF requires costly fine-tuning, demanding around 40 GPU days [4]. DiTFlow [31] offers a more efficient alternative by optimizing latents or positional embeddings at test time without model retraining. However, it incurs high computational costs due to its reliance on a full-size attention-based feature computation.

We extend DiTFlow's approach by updating only positional embeddings, avoiding latent space deviation and content leakage. Our method introduces motion-augmented rotary positional embeddings, warped via optical flow-derived displacements to embed motion cues (Section 4.1). We enhance this with flow-matching-guided optimization during early time steps, enabling stable and precise generation (Section 4.2). We further ensure spatiotemporal consistency through a Fourier phase regularization (Section 4.3). Unlike GWTF, our method requires no backbone training, drastically cutting computational costs. Compared to DiTFlow, it delivers higher efficiency and better motion quality. In addition, to evaluate motion alignment, we propose Fréchet Trajectory Distance (FTD) (Section 5.2). Our method outperforms recent approaches in qualitative and quantitative assessments.

Our contributions are:

- An efficient motion transfer, RoPECraft, that leverages motion-augmented rotary positional embeddings in a training-free setting, without requiring any backbone fine-tuning.
- A novel use of optical flow displacements to warp rotary positional embeddings, encoding spatial motion cues in attention calculations.
- A unified optimization strategy combining flow matching velocity prediction and phase constraint regularization to enhance motion accuracy and ensure temporal coherence.
- A new evaluation metric, Fréchet Trajectory Distance (FTD), for quantifying motion alignment between generated and reference videos.

Figure 2: Latent warping [4] without an expensive fine-tuning of the DiT fails (Column 2), and latent optimization is not adequate to recover the domain shift (Column 3). Our approach keeps the latent space intact, and performs successful motion transfer, all without model re-training (Column 4).

2 Related Work

Text-to-video models. Following the successful application of diffusion models in image generation [12, 30, 29, 34], efforts have been made to extend this approach to video generation. Initial efforts used U-Net-based architectures for this purpose, utilizing temporal modules [16] or inflated convolutions [1, 3, 20, 39] to transfer the image prior to the video domain. More recently, DiT pipelines [21, 44, 38, 25, 50, 28] have gained attention due to their superior capabilities in temporal modeling and enhanced quality. Motivated by these, we also adopt a DiT backbone.

Motion transfer. The goal of motion transfer is to synthesize videos whose dynamics match a reference clip while disentangling motion from appearance. Earlier methods injected explicit structure (masks or depth maps) [9, 43, 40]. Subsequent work learned dedicated motion embeddings and fed them to the generator [23, 22]. Recent approaches exploit the dense motion signals already present in backbone features [42, 15, 45, 14], or condition on trajectories extracted from the reference [48, 46, 41]. The current state of the art include Go With The Flow [4], which warps the initial noise with reference flow and fine-tunes the DiT on this prior, and DiTFlow [31], which derives displacement maps from cross-frame attention and updates either latents or positional embeddings. Building on DiTFlow's insight, we dynamically update RoPE to guide attention toward reference motion while keeping the backbone frozen. Unlike prior work, we initialize RoPE with our motion-augmentation algorithm and regularizers, enabling fast and accurate motion transfer, without requiring model fine-tuning [4], inversion [42, 45, 14], masks [14], and high GPU memory during tuning [31].

Positional embeddings in vision transformers. Transformers lack inherent order awareness, so Vision Transformers (ViTs) [11] rely on positional embeddings to encode spatial relationships among patches. In early ViT's fixed sinusoidal or learnable absolute embeddings were used [11, 8]. These failed to generalize across varying input resolutions or sequence lengths in videos [8, 13]. Rotary Position Embedding (RoPE) overcomes these issues by rotating query and key values according to patch positions, thereby capturing relative spatial or temporal relationships [35] . Originally successful in language models [35, 2], RoPE has since been adapted to vision models [27, 17, 29, 21, 44, 38]. Building on these advances, our method updates RoPE embeddings on the fly during generation.

3 Preliminaries

3.1 Flow matching

Flow Matching (FM) is a generative modeling approach that learns a deterministic, time-dependent velocity field to transform a simple base distribution into a complex target distribution [26]. Unlike diffusion models, which reverse stochastic processes [19], FM minimizes the discrepancy between the model velocity $v_\theta(t, x)$ and a target velocity $u_t(x)$ derived from the continuity equation:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,x\sim p_t} \|v_\theta(t, x) - u_t(x)\|^2 \tag{1}$$

This ensures mass-preserving transport along a predefined probability path p_t, enabling more efficient training and sampling than traditional diffusion models.

3.2 Rotary position embeddings (RoPE)

RoPE [35] encodes position by rotating query and key values $\mathbf{x}$ in the complex plane, enabling the model to capture relative positional relationships. Given a token at position m with vector $\mathbf{x}_m \in \mathbb{R}^d$, the vector is split into $d/2$ pairs. Each pair $(\mathbf{x}_m^{(2i-1)}, \mathbf{x}_m^{(2i)})$ is interpreted as a complex number $\mathbf{z}_m^{(i)} = \mathbf{x}_m^{(2i-1)} + j\,\mathbf{x}_m^{(2i)}$, and RoPE applies the rotation $\mathbf{z}_m^{(i)} \cdot \Phi_{m,i}$, where $\Phi_{m,i} = e^{jm\,\theta^{-2i/d}}$ is constructed by Algorithm 1, and θ is the base frequency. This operation embeds position into the phase of each frequency component, enabling self-attention to capture relative positions through inner products of queries and keys. Since attention patterns can control motion, we leverage RoPE heavily for our motion transfer task.

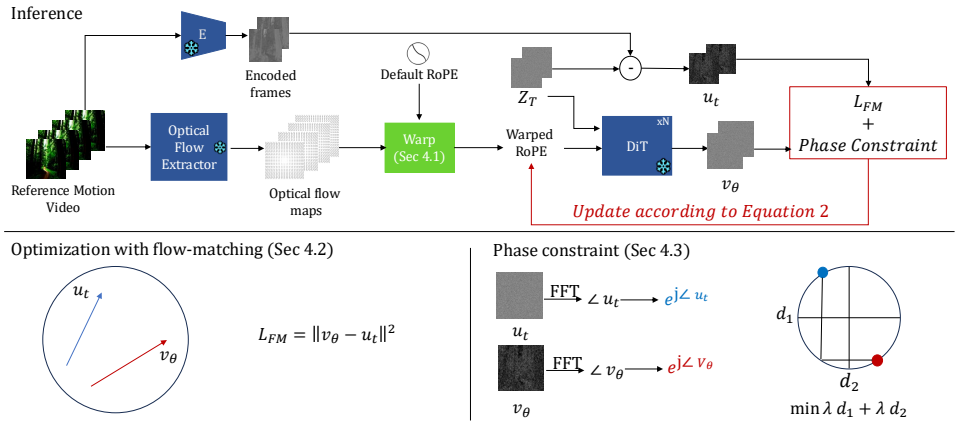


Figure 3: Visual description of our proposed pipeline inference and RoPE optimization approach.

4 Methodology

This section thoroughly explains the proposed components of our motion transfer method. The overall architecture is given in Fig. 3. We augment the RoPE tensors via optical flow maps (Section 4.1), and optimize them during the generation process (Section 4.2) with additional constraints (Section 4.3).

4.1 Motion-augmented RoPE

Algorithm 1 Default 1D RoPE, expanded to 3D

Input: Base frequency $\theta \in \mathbb{R}_{>0}$
Embedding dims $D_t, D_h, D_w \in \mathbb{N}$
Sequence lengths $S_t, S_h, S_w \in \mathbb{N}$
for each $k \in \{t, h, w\}$ **do**
  $\mathbf{p} = [0, 1, \ldots, S_k - 1]^{\mathrm{T}}$ ▷ $\in \mathbb{R}^{S_k}$
  $\mathbf{d} = [0, 1, \ldots, D_k/2 - 1]^{\mathrm{T}}$ ▷ $\in \mathbb{R}^{D_k/2}$
  $\mathbf{f} = \theta^{-2\mathbf{d}/D_k}$ ▷ $\in \mathbb{R}^{D_k/2}$
  $\mathbf{\Phi}_k = e^{j\mathbf{p}\mathbf{f}^{\mathrm{T}}}$ ▷ $\in \mathbb{C}^{S_k \times (D_k/2)}$
  $\mathbf{\Phi}_k = \texttt{expand}(\mathbf{\Phi}_k)$ ▷ $\in \mathbb{C}^{S_t \times S_h \times S_w \times (D_k/2)}$
end for
$\mathbf{\Phi} = \texttt{concat}(\mathbf{\Phi}_{t,h,w})$ ▷ $\in \mathbb{C}^{S_t \times S_h \times S_w \times (D/2)}$
$\mathbf{\Phi} = \texttt{flatten}(\mathbf{\Phi})$ ▷ $\in \mathbb{C}^{1 \times 1 \times (S_t S_h S_w) \times (D/2)}$

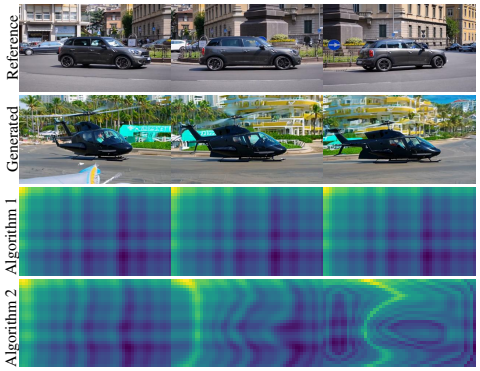


Figure 4: Comparison of the generations of default and motion-augmented RoPE.

The default RoPE algorithm used in video-DiTs is presented in Algorithm 1, where standard 1D RoPE is independently applied along the temporal (t), height (h), and width (w) dimensions to produce the respective components $\mathbf{\Phi}_t$, $\mathbf{\Phi}_h$, and $\mathbf{\Phi}_w$. These are then combined to form the full 3D positional encodings $\mathbf{\Phi}$, where each dimension $k \in \{t, h, w\}$ in $\mathbf{\Phi}_k$ is expanded (repeated) in the other two dimensions, $\{t, h, w\} \setminus k$. However, as previously discussed, our insight is that this formulation can be altered significantly with motion signals. Specifically, by having unique, motion-augmented 1D RoPEs for h and w components, we allow the attention mechanism during the generation process to better understand which spatial patches should attend to one another.

Our proposed procedure is detailed in Algorithm 2. For each row and column, we use the processed motion signals $\mathbf{h}_{\text{flow}}$ and $\mathbf{w}_{\text{flow}}$ to adjust the positional indices $\mathbf{p}$ in the complex exponential $\mathbf{\Phi} = \exp(j\mathbf{p}\mathbf{f}^{\mathrm{T}})$. Unlike Algorithm 1, where $\mathbf{\Phi}_h$ is fixed across all rows, and $\mathbf{\Phi}_w$ across all columns, Algorithm 2 introduces variation based on the motion signals. This way, we can construct unique embeddings for each spatial row and column for $\mathbf{\Phi}_h$ and $\mathbf{\Phi}_w$, respectively, providing a better initial condition for a motion-guided generation. We leave the temporal component $\mathbf{\Phi}_t$ unmodified, as altering it often introduces decoding artifacts without significant benefit.

Fig. 4 compares the default (Row 3) and modified (Row 4) embeddings visually. It can be observed that Algorithm 2 warps RoPE tensors in the motion direction (Row 1), whose effects are reflected on the generation (Row 2). Fig. 5 demonstrates the effectiveness of our approach across various prompts, which transfers coarse input motion directly into the generated videos.

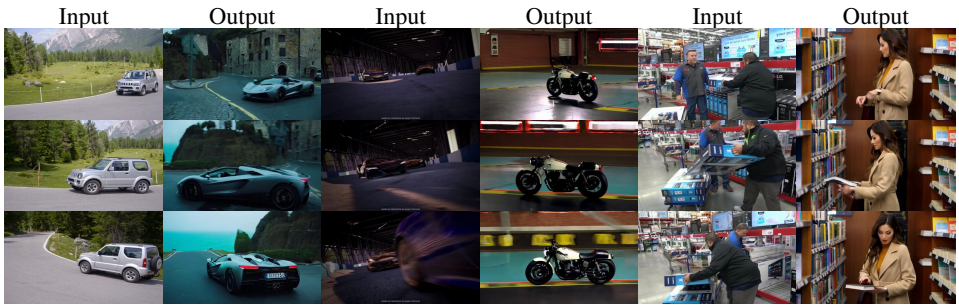


Figure 5: Qualitative results of motion-augmented RoPE described in Algorithm 2.

However, relying solely on the modification introduced in Algorithm 2 can yield suboptimal results. For example, while the overall motion may be correct, subjects sometimes face the opposite direction (Column 4) or fail to accurately follow challanging trajectories (Column 6). To address these limitations, we introduce a brief optimization step over the motion-augmented RoPE tensors during generation, which will be described in Section 4.2.

4.2 Optimization with flow-matching

Algorithm 2 Motion-augmented RoPE

Input: Base frequency $\theta \in \mathbb{R}_{>0}$,
Embedding dims $D_t, D_h, D_w \in \mathbb{N}$
Sequence lengths $S_t, S_h, S_w \in \mathbb{N}$
Optical flows $\mathbf{u}, \mathbf{v}$ $\triangleright \in \mathbb{R}^{2\times S_t \times H \times W}$

$\mathbf{u}, \mathbf{v} = \texttt{downsample}(\mathbf{u}, \mathbf{v})$ $\triangleright \in \mathbb{R}^{2\times S_t \times S_h \times S_w}$
$\mathbf{h}_{\text{flow}}, \mathbf{w}_{\text{flow}} = \texttt{cumsum}(\mathbf{u}, \mathbf{v})$
$\mathbf{h}_{\text{flow}} = \texttt{flatten}(\mathbf{h}_{\text{flow}})$ $\triangleright \in \mathbb{R}^{(S_t \times S_w)\times S_h}$
$\mathbf{w}_{\text{flow}} = \texttt{flatten}(\mathbf{w}_{\text{flow}})$ $\triangleright \in \mathbb{R}^{(S_t \times S_h)\times S_w}$

$\mathbf{f}_h = \theta^{-2[0,1,\ldots,D_h/2-1]^{\mathrm{T}}/D_h}$ $\triangleright \in \mathbb{R}^{D_h/2}$
for each r in $[0, 1, \ldots, S_t \times S_w]$ **do**
  $\mathbf{p} = [0, 1, \ldots, S_h - 1]^{\mathrm{T}} + \mathbf{h}_{\text{flow}}[r]$ $\triangleright \in \mathbb{R}^{S_h}$
  $\mathbf{\Phi}_h[r] = e^{j\mathbf{p}\mathbf{f}_h^{\mathrm{T}}}$ $\triangleright \in \mathbb{C}^{S_h \times (D_h/2)}$
end for
$\mathbf{\Phi}_h = \texttt{reorder}(\mathbf{\Phi}_h)$ $\triangleright \in \mathbb{C}^{S_t \times S_h \times S_w \times (D_h/2)}$

$\mathbf{f}_w = \theta^{-2[0,1,\ldots,D_w/2-1]^{\mathrm{T}}/D_w}$
for each c in $[0, 1, \ldots, S_t \times S_h]$ **do**
  $\mathbf{p} = [0, 1, \ldots, S_w - 1]^{\mathrm{T}} + \mathbf{w}_{\text{flow}}[c]$
  $\mathbf{\Phi}_w[c] = e^{j\mathbf{p}\mathbf{f}_w^{\mathrm{T}}}$
end for
$\mathbf{\Phi}_w = \texttt{reorder}(\mathbf{\Phi}_w)$

$\mathbf{p} = [0, \ldots, S_t - 1]^{\mathrm{T}}$
$\mathbf{f} = \theta^{-2[0,\ldots,D_t/2-1]^{\mathrm{T}}/D_t}$
$\mathbf{\Phi}_t = \texttt{expand}(e^{j\mathbf{p}\mathbf{f}^{\mathrm{T}}})$

$\mathbf{\Phi} = \texttt{flatten}(\texttt{concat}(\mathbf{\Phi}_{t,h,w}))$

To refine Algorithm 2, we apply a brief optimization on rotary embeddings during early generation steps. Using Eq. (1), we align the generated velocity $v_\theta(t, x_t)$ with the target velocity $u_t(x) = \sigma_t^{-1}(x_t - \mathbf{v})$, where x_t is the current latent in time step t, $\mathbf{v}$ is latent reference video, and σ is the scheduler sigma.

Fig. 6 illustrates the effectiveness of both optimization and the motion-augmented RoPE initial condition. In Column 1, the subject moves away from the camera, while in Column 2, the subject moves from left to right. The motion-augmented RoPE approach (Columns 3–4) successfully captures the general movements. However, in the second sample, it incorrectly renders the motorbike facing backward. When optimization is performed without a dedicated initial condition (Algorithm 1), the subject placement improves, but issues arise in motion direction (Column 6), and visual artifacts appear (Column 5, Row 2). In contrast, initializing with Algorithm 2 yields the best results across the samples. This approach reduces artifacts, corrects subject orientation and trajectories.

4.3 Phase constraints

Flow matching optimization produces strong results, as shown in Fig. 6, but we occasionally observe duplicated subjects when adjusting the orientation, position, or motion of the moving subjects. To address this, we build on insights from prior work [47] and analyze the Fourier transform of our signals. Since linear displacements in the spatial domain cause a phase shift in frequency domain,

Figure 6: Qualitative results on optimization, with identical seeds across different experiments.

a Fourier property closely tied to motion transfer, we add a phase constraint to the flow matching objective to guide the model toward more accurate and consistent spatiotemporal alignment.

Specifically, we take the Fourier transform of the target velocity u_t along spatio-temporal dimension, to get $\mathcal{F}(u_t) = \mathbf{U}_t = |\mathbf{U_t}| \exp\{j\angle \mathbf{U_t}\}$, where $|\mathbf{U_t}| = \{\mathfrak{Re}(\mathbf{U_t})^2 + \mathfrak{Im}(\mathbf{U_t})^2\}^{1/2}$ is the magnitude, and $\angle \mathbf{U_t} = \arctan\{\mathfrak{Im}(\mathbf{U_t})/\mathfrak{Re}(\mathbf{U_t})\}$ is the phase. We perform the same transform to the DiT output v_θ, and add the phase constraint as a $\mathcal{L}_1$ regularizer to the main optimization objective. We represent the phase on the unit circle ($\exp\{j\angle\mathcal{F}(\cdot)\}$) to make them continuous and differentiable everywhere, since the original mapping $\angle\mathcal{F}(\cdot)$ contains jump discontinuities at $\pm\pi$ due to being bounded by $(-\pi, \pi]$. More clearly, we represent $\exp\{j\angle\mathcal{F}(\cdot)\} = \cos(\angle\mathcal{F}(\cdot)) + j\sin(\angle\mathcal{F}(\cdot))$ and perform the phase-consistency loss in two parts. The final optimization objective is given in Eq. (2):

$$\min \mathcal{L}_{\text{FM}}(u_t, v_\theta) + \lambda \| \cos\angle\mathcal{F}(u_t) - \cos\angle\mathcal{F}(v_\theta)\| + \lambda\| \sin\angle\mathcal{F}(u_t) - \sin\angle\mathcal{F}(v_\theta)\|, \qquad (2)$$

where λ is the hyperparameter.

Fig. 7 reveals the effect of phase constraints, fixing duplicate generations and artifacts. Additional experiments regarding Fourier components are given in Supplementary.

Figure 7: Qualitative results on phase constraints, with identical seeds across different experiments.

5 Experimental Results

5.1 Metrics and baselines

We compare our method with the recently published motion transfer methods [4, 14, 31, 45, 42]. For evaluation, we use videos from the DAVIS dataset [32]. We generate 4 diverse prompts per DAVIS video via ShareGPT4V [7] and LLAMA 3.2 [36], which are detailed in the Supplementary.

For the evaluation, we use content-debiased Fréchet Video Distance (CD-FVD) [5] for evaluating fidelity, CLIP similarity [18] for evaluating frame-wise prompt fidelity using ViT B/32 model [33], and Motion Fidelity (MF) [45] along with our proposed metric Fréchet Trajectory Distance (FTD) for evaluating the motion alignment between the generated and the ground truth reference motion video. For assessing the motion of the foreground object as well as camera motion, we use FTD by only sampling from the foreground object mask region, and sampling from both the foreground and background mask region. For MF and FTD, the trajectories are obtained using Co-Tracker3 [24].

For video synthesis, we use Wan2.1-1.3B [38] as the backbone of our method. To obtain a fair assessment, similar to the approach used in DiTFlow [31], we adapt MOFT [42], SMM [45], DitFlow [31] and ConMo [14] to Wan2.1. For the methods that require DDIM inversion [42, 45, 14],

we perform KV-injection from reference video latents similar to [31]. We evaluate GWTF using their CogVideoX-2B [21] checkpoint. The hyper parameters are detailed in Supplementary.

5.2 Fréchet Trajectory Distance

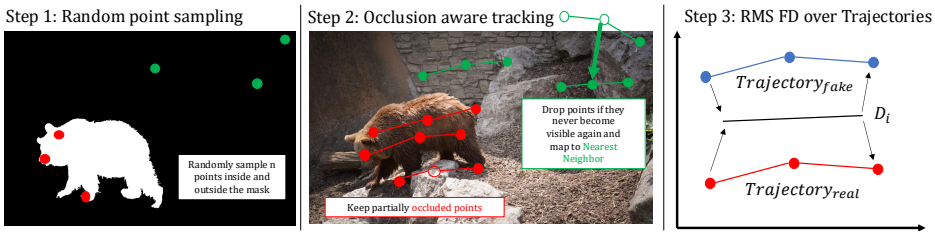


Figure 8: **Fréchet Trajectory Distance (FTD). 1)** Sample n foreground (**red**) and n background (**green**) seeds on the first frame. **2)** Track each seed with an occlusion-aware filler: copy the nearest visible neighbor while occluded and discard tracks that never re-appear. **3)** Measure the RMS Fréchet distance between generated (**fake**) and reference (**real**) tracks.

Discrete Fréchet Distance. Let $\mathbf{x}_{i,t} \in \mathbb{R}^2$ be the 2D image coordinate of the i^{th} point at frame t ($1 \le t \le T$). We denote the reference and generated trajectories by $\mathcal{T}_i^{\text{real}} = \{\mathbf{x}_{i,1}^{\text{real}}, \ldots, \mathbf{x}_{i,T}^{\text{real}}\}$ and $\mathcal{T}_i^{\text{fake}} = \{\mathbf{x}_{i,1}^{\text{fake}}, \ldots, \mathbf{x}_{i,T}^{\text{fake}}\}$, respectively. Then, the discrete Fréchet distance is defined as:

$$D_F\big(\mathcal{T}_i^{\text{real}}, \mathcal{T}_i^{\text{fake}}\big) = \min_{\sigma,\tau:\{1,\ldots,L\}\to\{1,\ldots,T\}} \max_{k=1,\ldots,L} \big\|\mathbf{x}_{i,\sigma(k)}^{\text{real}} - \mathbf{x}_{i,\tau(k)}^{\text{fake}}\big\|_2, \tag{3}$$

where L is the length of the common re-parameterization, and (σ, τ) are non-decreasing index maps that allow each curve to pause or advance but never step backwards. The inner $\max$ takes the worst spatial gap along a particular pairing of frames, while the outer $\min$ selects the pairing that makes this worst gap as small as possible. Consequently, D_F is the minimal worst-case deviation between the trajectories after they are aligned in time as favorably as the monotone constraint permits.

Fréchet Trajectory Distance (FTD). We utilize Eq. (3) in our proposed FTD metric, along with several tricks to obtain meaningful trajectories $\mathcal{T}$. Fig. 8 explains our procedure thoroughly. From the first frame, we uniformly select n points inside the binary foreground mask $\mathcal{M}_0$, and n outside to capture both object (red) and background (green) motion. Then, the occlusion-aware tracker [24] generates tracks starting with the initial $2n$ points. Since some points may go out of bounds or get occluded as the video progresses, we reassign them by copying to their nearest visible neighbor to maintain trajectory continuity, and drop a track entirely if the associated point never reappears. Before computing distances, all coordinates are normalized by the frame width W and height H, making the metric resolution-invariant. The procedure yields N valid, temporally coherent pairs $\{\mathcal{T}_i^{\text{real}}, \mathcal{T}_i^{\text{fake}}\}_{i=1}^N$. Utilizing the pairs, we calculate root-mean-square Fréchet distance, FTD $= (N^{-1}\Sigma_{i=1}^N D_F^2(\mathcal{T}_i^{\text{real}}, \mathcal{T}_i^{\text{fake}}))^{0.5}$. For calculating D_F, we utilize [10].

Comparison with Motion Fidelity [45]. The Motion Fidelity (MF) metric [45] computes cosine similarity between frame-to-frame displacements on a fixed grid, averaging best matches. However, it ignores path shape, magnitude, and occlusions, and can report high scores even when trajectories diverge. In contrast, our Fréchet Trajectory Distance (FTD) drops unreliable tracks, focuses on relevant regions, and measures curve distance using discrete Fréchet distance, making it more robust to missing data and outliers. As shown in Table 1, MF also exhibits much higher standard deviation, highlighting its instability compared to FTD.

To further illustrate the behavioral difference between the two metrics, we present two controlled toy examples in Fig. 9. In Case 1, the generated trajectory follows the same path as the reference but with a smaller motion magnitude, while in Case 2, both trajectories share identical motion directions yet are spatially offset. Because the Motion Fidelity (MF) metric normalizes motion vectors to unit length and computes directional cosine similarity over a fixed grid, it reports perfect similarity (MF=1.0) in both cases. This normalization causes MF to ignore motion speed, path magnitude, and spatial alignment, effectively making it translation-invariant but insensitive to geometric deviation. Moreover, MF can be artificially inflated by nearest-neighbor matches anywhere in the frame, as it measures

average directional alignment between motion vectors rather than actual trajectory correspondence. In contrast, the Fréchet Trajectory Distance (FTD) evaluates the geometric trajectory consistency over time by measuring the curve distance between corresponding paths. FTD therefore penalizes drift, speed changes, and path-shape differences directly. FTD increases proportionally with geometric and temporal discrepancies (FTD=1.414 and 1.0 for Cases 1 and 2, respectively), demonstrating its discriminative power and perceptual robustness.

5.3 Qualitative evaluation

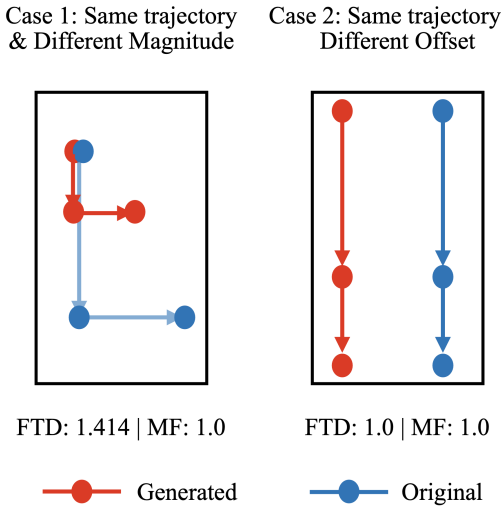


Figure 9: Illustration of the behavioral difference between Motion Fidelity (MF) and Fréchet Trajectory Distance (FTD) across two controlled toy cases. In both cases, the generated trajectories (red) differ from the original trajectories (blue) either in magnitude (Case 1) or in spatial offset (Case 2). Although both pairs exhibit substantial geometric deviation, MF remains artificially high because it measures only the directional alignment of motion vectors, discarding scale and positional information. In contrast, FTD penalizes such discrepancies by accounting for the full geometric path similarity, thereby providing a more faithful measure of motion consistency. It is important to note that FTD evaluates distance, hence smaller values mean better motion alignment where MF evaluates direct motion alignment, hence larger values is better.

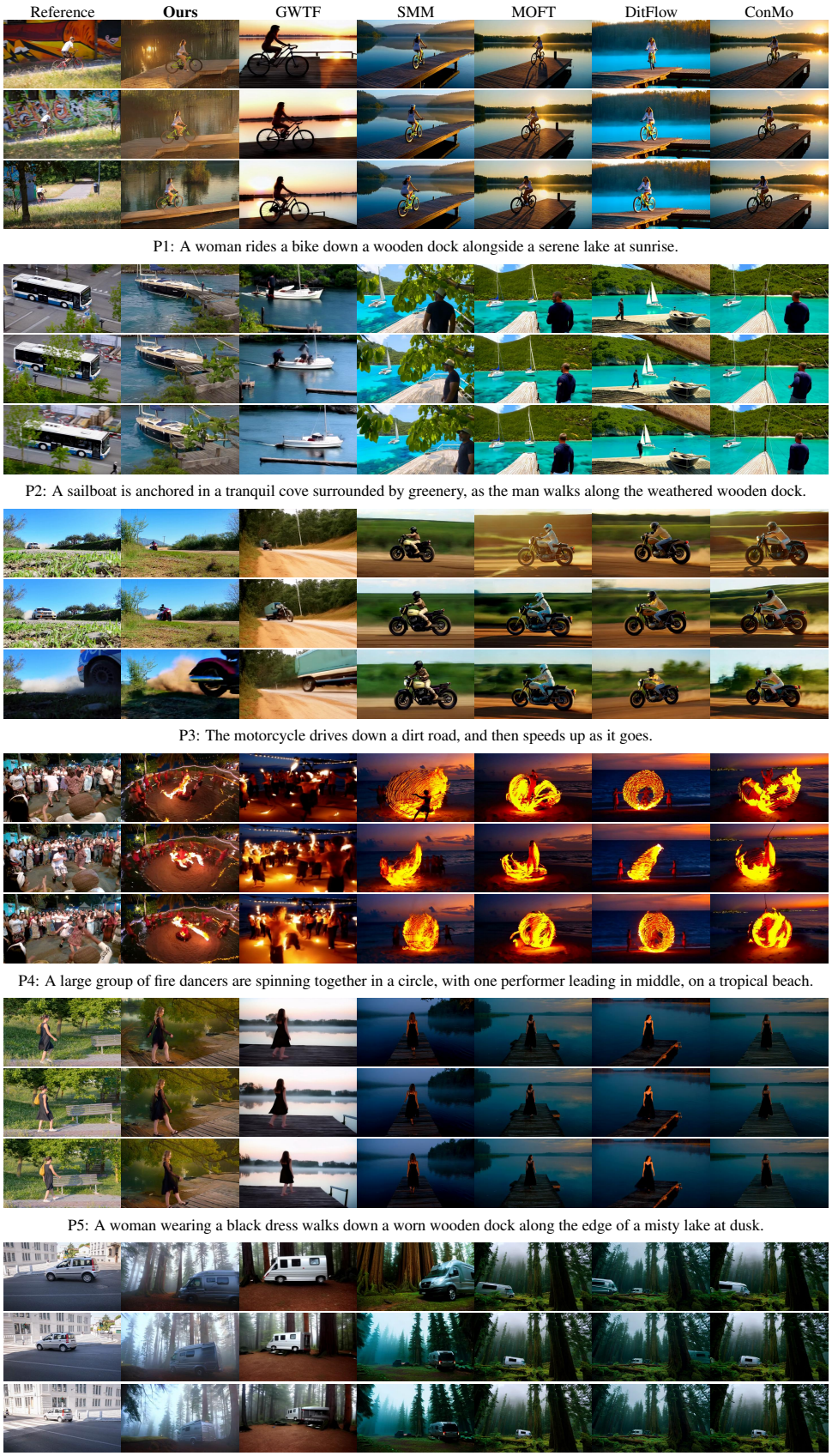
Fig. 10 provides a visual comparison of the evaluated methods across diverse prompts and motion scenarios. Our approach consistently outperforms others in both trajectory direction and subject orientation. In P1, MOFT, DitFlow, ConMo, and SMM fail to capture the correct motion direction, although SMM maintains proper subject orientation. In P2, some methods struggle with prompt alignment, such as keeping the man stationary, and GWTF introduces noticeable artifacts. For more complex motions like P3 and P4, most methods do not use the reference motion effectively. While GWTF shows motion coherence, it often sacrifices from prompt alignment. For example, it merges a motorcycle with a truck in P3 and does not place the man walking on the wooden dock in P2. A similar issue appears in P6, where GWTF generates a distorted motorhome, and only SMM, GWTF, and our method reflect the reference motion correctly. In general, our method accurately captures both motion and subject across all examples. Additional results of our model on challenging videos such as videos with camera motion or multiple subjects can be seen in the Section A.6.

5.4 Quantitative evaluation

Table 1 compares our method with recent motion transfer baselines across five key metrics: Motion Fidelity (MF) [45], content-debiased Fréchet Video Distance (CD-FVD) [5], CLIP similarity [18], and our occlusion-aware FTD on foreground (FG) and all points (FG+BG).

Our method achieves the highest MF score (**0.5816**) and the lowest CD-FVD (**1284.58**), surpassing a strong baseline, GWTF [4], by +0.0103 (approximately +1.8%) and -200.6 (approximately -13.5%), respectively. It also attains the second-best CLIP similarity (**0.2350**), and ranks second on both FTD variants, **0.2644** for FG and **0.2584** for FG+BG, while outperforming all remaining competitors. In terms of runtime, our method runs at **109.231** $\pm$ **3.112 s**, comparable to SMM [45] (107.281 $\pm$ 4.562 s), DitFlow (RoPE) [31] (104.405 $\pm$ 2.056 s), DitFlow (Latent) [31] (105.126 $\pm$ 2.739 s), MOFT [42] (119.411 $\pm$ 3.847 s), and GWTF [4] (101.342 $\pm$ 3.337 s), while being significantly faster than ConMo [14] (150.666 $\pm$ 3.317 s). This demonstrates that our framework achieves a strong trade-off between computational efficiency and high-fidelity generation quality.

We also showcase quantitative scores in ablation studies, validating the effectiveness of our design. Specifically, Table 2 justifies our approach on motion-augmented RoPE, optimization procedure with flow-matching, and phase constraints. Table 3 elaborates on the selection of first t denoising steps in our optimization, and s number of optimization steps per t. We opted for $(t, s) = (10, 5)$ as we noticed that $s = 10$ decreases visual quality significantly.



P1: A woman rides a bike down a wooden dock alongside a serene lake at sunrise.

P2: A sailboat is anchored in a tranquil cove surrounded by greenery, as the man walks along the weathered wooden dock.

P3: The motorcycle drives down a dirt road, and then speeds up as it goes.

P4: A large group of fire dancers are spinning together in a circle, with one performer leading in middle, on a tropical beach.

P5: A woman wearing a black dress walks down a worn wooden dock along the edge of a misty lake at dusk.

P6: A silver motorhome pulling up to a campsite nestled among the towering redwoods of a misty forest.

Figure 10: Qualitative comparison of the methods with diverse prompts.

Table 1: Comparison of motion transfer methods across evaluation metrics. Best and second results are represented with *italic* and underlined, respectively.

Method	MF ↑	CD-FVD ↓	CLIP ↑	FTD (FG) ↓	FTD (FG+BG) ↓
GWTF [4]	0.5713±0.22	1485.23	*0.2378±0.04*	*0.2457±0.14*	*0.2308±0.10*
SMM [45]	0.4889±0.20	1600.33	0.2331±0.04	0.2882±0.15	0.3176±0.15
MOFT [42]	0.4606±0.20	1630.45	0.2311±0.04	0.2811±0.16	0.3057±0.14
DitFlow (latents) [31]	0.4832±0.20	1735.49	0.2339±0.04	0.2921±0.15	0.3135±0.12
DitFlow (RoPE) [31]	0.4500±0.18	1852.90	0.2345±0.04	0.2785±0.14	0.3019±0.13
ConMo [14]	0.4627±0.21	1680.78	0.2309±0.04	0.2769±0.15	0.3040±0.14
Ours	*0.5816±0.19*	*1284.58*	0.2350±0.04	0.2644±0.14	0.2584±0.13

Table 2: Ablation on motion-augmented RoPE and phase constraints.

Method	MF	CLIP	FTD
Alg.1 + opt.	0.7082	0.1560	0.2174
Alg.2 + opt.	0.7092	0.1650	0.2105
Alg.2 + opt. + phase	*0.7210*	*0.1656*	*0.2060*

Table 3: Ablation on hyperparameters.

(t, s)	MF	CD-FVD	CLIP	FTD
$5, 5$	0.5165	1437.99	0.1597	0.2901
$5, 10$	0.5523	1606.88	0.1663	0.2728
$10, 5$	0.5675	*1364.25*	*0.1664*	0.2633
$10, 10$	*0.6160*	1492.86	0.1572	*0.2573*

5.5 User study

We conducted a user study to evaluate (i) how well our proposed FTD and MF metrics align with human perception, and (ii) overall method quality. We randomly sampled 20 prompt-reference pairs from DAVIS and generated outputs for all competing methods. In the first part of the survey, participants selected the top-3 videos that best matched the reference motion. As shown in Table 4, FTD exhibits a noticeably stronger correlation with human motion judgments than MF. In the second part, users ranked their top-3 videos based on overall visual preference. The results in Table 5 show that our method is consistently preferred across all evaluated dimensions.

Table 4: Alignment of motion preference.

	FTD (%)	MF (%)
1st choice	**25**	11
2nd choice	**17**	17
3rd choice	**19**	11

Table 5: User preference study for visual quality.

Method	RPC	GWTF	SMM	MOFT	ConMo	DF-L	DF-R
1st	**30%**	10%	13%	19%	10%	12%	6%
2nd	**23%**	12%	14%	11%	13%	13%	14%
3rd	13%	11%	22%	13%	16%	13%	12%

These findings reveal two key outcomes: (1) FTD aligns better with human motion perception than MF, and (2) RoPECraft is consistently preferred over all baselines in overall quality.

6 Conclusion and Discussion

In this paper, we introduce RoPECraft, a training-free motion transfer method that manipulates rotary positional embeddings in diffusion transformers. By combining motion-augmented RoPE tensors, flow-matching-based optimization, and phase-based regularization, RoPECraft achieves high-quality performance across multiple benchmarks, and produces high-quality motion transfer results.

For the future work, the motion-augmented RoPE framework can be extended to handle more challenging cases, such as handling motion with extreme occlusion, and better high-frequency details in the generated videos. In addition, the pipeline can be extended to controllable video editing.

We discuss limitations and broader impacts in the Supplementary Material.

Acknowledgements. We acknowledge EuroHPC Joint Undertaking for awarding the project ID EHPC-AI-2024A02-031 access to Leonardo at CINECA, Italy. We also acknowledge Fal.ai for granting GPU access.

References

[1] 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: ACM SIGGRAPH (2024)

[2] Barbero, F., Vitvitskyi, A., Perivolaropoulos, C., Pascanu, R., Veličković, P.: Round and round we go! what makes rotary positional encodings useful? In: Proceedings of the International Conference on Learning Representations. ICLR (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., Ryoo, M., Debevec, P., Yu, N.: Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR (2025)

[5] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR (2017)

[6] Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR (2024)

[7] Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: Proceedings of the European Conference on Computer Vision. ECCV (2024)

[8] Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. In: Proceedings of the International Conference on Learning Representations. ICLR (2023)

[9] 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)

[10] Denaxas, S., Pikoula, M.: spiros/discrete_frechet: meerkat stable release, `https://github.com/spiros/discrete_frechet`

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations. ICLR (2020)

[12] 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: Proceedings of the International Conference on Machine Learning. ICML (2024)

[13] Fan, Q., You, Q., Han, X., Liu, Y., Tao, Y., Huang, H., He, R., Yang, H.: Vitar: Vision transformer with any resolution. arXiv preprint arXiv:2403.18361 (2024)

[14] 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR (2025)

[15] Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. In: Proceedings of the International Conference on Learning Representations. ICLR (2024)

[16] 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: Proceedings of the International Conference on Learning Representations. ICLR (2024)

[17] Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: Proceedings of the European Conference on Computer Vision. ECCV (2024)

[18] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP (2021)

[19] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the International Conference on Neural Information Processing Systems. NeurIPS (2020)

[20] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: Proceedings of the International Conference on Neural Information Processing Systems. NeurIPS (2022)

[21] Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In: Proceedings of the International Conference on Learning Representations. ICLR (2023)

[22] Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR (2024)

[23] Kansy, M., Naruniec, J., Schroers, C., Gross, M., Weber, R.M.: Reenact anything: Semantic video motion transfer using motion-textual inversion. arXiv preprint arXiv:2408.00458 (2024)

[24] Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. arXiv preprint arXiv:2410.11831 (2024)

[25] 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)

[26] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: Proceedings of the International Conference on Learning Representations. ICLR (2023)

[27] Liu, Z., Guo, L., Tang, Y., Cai, J., Ma, K., Chen, X., Liu, J.: Vrope: Rotary position embedding for video large language models. arXiv preprint arXiv:2502.11664 (2025)

[28] Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. Transactions on Machine Learning Research (2025)

[29] Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. CVPR (2023)

[30] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: Proceedings of the International Conference on Learning Representations. ICLR (2024)

[31] Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR (2025)

[32] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)

[33] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. ICML (2021)

[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. CVPR (2022)

[35] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

[36] Team, L.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783

[37] Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Proceedings of the European Conference on Computer Vision. ECCV (2020)

[38] Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

[39] Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)

[40] Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. In: Proceedings of the International Conference on Neural Information Processing Systems. NeurIPS (2023)

[41] 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)

[42] Xiao, Z., Zhou, Y., Yang, S., Pan, X.: Video diffusion models are training-free motion interpreter and controller. In: Proceedings of the International Conference on Neural Information Processing Systems. NeurIPS (2024)

[43] 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 (2024)

[44] 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: Proceedings of the International Conference on Learning Representations. ICLR (2025)

[45] 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. CVPR (2024)

[46] 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)

[47] Yuan, Y., Guo, Y., Wang, C., Zhang, W., Xu, H., Zhang, L.: Freqprior: Improving video diffusion models with frequency filtering gaussian noise. In: Proceedings of the International Conference on Learning Representations. ICLR (2025)

[48] 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 IEEE/CVF international conference on computer vision. CVPR (2025)

[49] Zheng, G., Li, T., Zhou, X., Li, X.: Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. arXiv preprint arXiv:2504.08212 (2025)

[50] Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions are stated in abstract and summarized at the of Section 1 in bullet points.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We have discussed the limitations of our work and gave relevant examples in Section A.5

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not include any theoretical result hence there is no proof given.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All of the hyperparameters used in our experiments are given in Section A.3

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code for both model and proposed metric is provided at `https://github.com/berkegokmen1/RoPECraft`

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have discussed how we test our competitors and our baseline, which dataset we have used, type of hyperparameters and optimizer utilized in Section A.3

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have provided error bars in our comparison with other pipelines which can be seen in Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The type of compute resources utilize during our experiments is given in Section A.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have read and obey the NeurIPS Code of Ethics in full.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed our broader impacts in Section A.5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our model's backbone Wan 2.1 has internal safety checker for NSFW content baked in https://github.com/Wan-Video/Wan2.1/issues/190.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA] .

Justification: This paper does not provide any new assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: This paper does not propose any new assets as a result, this question is not applicable.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We have utilized LLMs for generating captions of the videos of our dataset and expanding these captions into more diverse prompts. This procedure is explained in detail at Section A.2

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Ablation on Fourier Features

We ablate the Fourier components for additional constraints on the flow-matching objective on our optimization stage. For the objective $\min \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{c}}$, we ablate two regularizer for magnitude and phase, Eq. (4) and Eq. (5), respectively,

$$\mathcal{L}_{\text{c}} = \lambda \left\| |\mathcal{F}(u_t)| - |\mathcal{F}(v_\theta)| \right\|_1 \tag{4}$$

$$\mathcal{L}_{\text{c}} = \lambda \| \cos \angle \mathcal{F}(u_t) - \cos \angle \mathcal{F}(v_\theta) \| + \lambda \|_1 \sin \angle \mathcal{F}(u_t) - \sin \angle \mathcal{F}(v_\theta) \|_1, \tag{5}$$

where u_t denotes the target velocity, and v_θ is the generated velocity output from the transformer at time step t. As shown in Fig. 11, the phase-based constraint proves more effective than the magnitude-based one, supporting the discussion in the main paper. λ is chosen as 1.0 across all experiments, as sweeping λ did not result in significant changes.

P1: A woman rides a bike down a wooden dock alongside a serene lake at sunrise.
P2: A silver motorhome pulling up to a campsite nestled among the towering redwoods of a misty forest.

Figure 11: Qualitative comparison of using magnitude and phase constraints. First column shows the video associated with P1, whereas the second column is with P2.

A.2 Prompts

We extract prompts from DAVIS [32] videos by using ShareGPT4V [7]. For generating diverse prompts, we utilize LLAMA 3.2 [36]. We construct different `system_prompts` for changing the object (1), and environment (2). We also provide a paraphrased prompt for reconstruction (3). The system prompts are listed below:

```
system_prompt_1 = Answer with a single sentence.  You will receive a
single-sentence or multi-sentence video prompt.  Replace its main subject
(the actor or object performing the action) with a new, physically
plausible subject while leaving the action, environment, camera movements,
and style intact.  The new subject must be realistic in the described
scenario (e.g., a golden retriever a border collie; a sports car a vintage
motorcycle).  Return only the fully rewritten prompt-no explanations, no
bullet points.  Keep the scene information in the prompt the same.  Do
not just change the gender etc.  like man-woman or woman-man.  Change the
entity class, do not just replace person with person.
```

```
system_prompt_2 = Answer with a single sentence.  You will receive a
video prompt.  Keep the subject(s) and their actions exactly the same, but
relocate the scene or setting to a coherent, vivid new environment.  Ensure
lighting, weather, and background details match the new setting and remain
physically reasonable.  Output the updated prompt text and nothing else.
```

```
system_prompt_3 = Answer with a single sentence.  Do not alter the subject,
action, or scene.  Simply rephrase the text so it is stylistically
```

`different (synonyms, varied sentence structure) while preserving every factual detail. Return the single paraphrased prompt-no commentary, no headings.`

For each original prompt from ShareGPT4V [7], we apply these system prompts to generate its corresponding response. The motion prompts utilized in the main paper figures are listed below:

- `A sleek, black helicopter is seen around a bustling beach side promenade, passing by a seaside resort building.`
- `A sleek, silver sports car is navigating through the foggy streets of an Italian Renaissance-era town perched on the edge of a rugged cliff overlooking the turquoise Mediterranean Sea.`
- `The video shows a vintage motorcycle driving down a track in a garage.`
- `A woman wearing a beige coat is seen browsing in a bookstore, examining a shelf and then selecting a book from the stack.`
- `A man is seen sprinting across a deserted beach at sunset, his feet pounding against the wet sand.`
- `A man on a motorcycle is seen riding down a coastal highway with rugged cliffs and rocky outcroppings lining the edge of the ocean, as sunlight catches the spray of the waves and casts a misty veil over the scene.`
- `A backpacker is seen walking on a rocky terrain with mountains in the background.`
- `A white van is seen driving down a street with a building in the background.`
- `A goose walks on grass and then flies over a river.`

A.3 Hyperparameters and Computational Requirements

Hyperparameters. For video synthesis, we adopt Wan2.1-1.3B [38] as the backbone of our method. Since DiTFlow [31] was originally proposed on CogVideoX [21], and MOFT [42], SMM [45], and ConMo [14] were developed on UNet-based architectures, we re-implemented all these baselines using the same Wan2.1-1.3B backbone for a fairer comparison. For methods that require DDIM inversion [42, 45, 14], we applied key-value (KV) injection into all transformer blocks during the first t denoising steps, following the strategy used in DiTFlow.

We conducted extensive experiments to determine optimal hyperparameters for each method in the Wan2.1 framework. The hyperparameters are defined as follows: learning rate (l), transformer block index for motion feature extraction (b), number of optimization steps (s), number of early denoising steps used for optimization (t, out of 50 total steps), AMF attention temperature for DitFlow (d), and mask fusion weight for ConMo (w). We utilize Adam optimizer across all methods, with their default β parameters.

1. **DiTFlow** [31]: $l = 1 \times 10^{-4}$, $b = 10$, $s = 10$, $t = 5$, $d = 2.0$
2. **MOFT** [42]: $l = 1 \times 10^{-4}$, $b = 10$, $s = 10$, $t = 5$
3. **SMM** [45]: $l = 1 \times 10^{-4}$, $b = 5$, $s = 10$, $t = 5$
4. **ConMo** [14]: $l = 1 \times 10^{-4}$, $b = 20$, $s = 5$, $t = 10$, $w = 0.5$
5. **Ours**: $l = 1 \times 10^{-4}$, $s = 5$, $t = 10$

For ConMo, we used ground-truth DAVIS masks for the reference videos. We do not extract motion cues from internal layers of the transformer, hence b is not applicable for our case.

Due to limited computational resources, we used the original CogVideoX weights provided by the authors for GWTF [4], as training the full Wan2.1 pipeline from scratch was infeasible. To align more closely with Wan's 1.3B parameter scale during evaluation, we used the 2B checkpoint of GWTF.

Computational Requirements. We run all the models on a shared cluster, with compute nodes equipped with $4\times$ NVIDIA A100 64GB.

A.4 Fréchet Trajectory Distance

We provide our Fréchet Trajectory Distance pseudocode in Listing 1, where the `frechetdist` is calculated by using [10] and `cotracker3` by [24].

Listing 1 Fréchet Trajectory Distance implementation.

```
import frechetdist
def fill_and_drop(track, vis):
    filled = track.clone()
    N, F, _ = filled.shape
    for t in range(1, F):
        inv_idx = (~vis[:, t]).nonzero(as_tuple=False).view(-1)
        vis_idx = vis[:, t].nonzero(as_tuple=False).view(-1)
        if inv_idx.numel() and vis_idx.numel():
            prev_pts = filled[inv_idx, t - 1]
            curr_pts = filled[vis_idx, t]
            d = distance_matrix(prev_pts, curr_pts)
            filled[inv_idx, t] = curr_pts[d.argmin(dim=1)]
        else:
            filled[:, t] = filled[:, t - 1]
    dropped = (~vis[:, 1:].any(dim=1)).nonzero(as_tuple=False).view(-1)
    return filled, dropped

def compare_trajectory_consistency(cotracker3 , video1, video2, mask,
                                   n_points=100, use_fg_mask_only=False):
    _, T, C, H, W = video1.shape
    if use_fg_mask_only:
        queries = sample_points_inside_mask_randomly(mask, n_points)
    else:
        queries = sample_points_from_mask_randomly(mask, fg=n_points//2,
        ↪  bg=n_points//2)
    tracks = []
    drops = []
    for vid in (video1, video2):
        pts, vis = cotracker3(vid, queries=queries)
        pts, drop = fill_and_drop(pts[0], vis[0])
        pts[...,0] /= W
        pts[...,1] /= H
        tracks.append(pts)
        drops.append(drop)
    sq = []
    for i in range(tracks[0].shape[1]):
        if i in drops[0] or i in drops[1]:
            continue
        P = tracks[0]
        Q = tracks[1]
        fd = frechetdist(P, Q)
        sq.append(fd*fd)
    return sqrt(mean(sq))
```

A.5 Limitations and Broader Impacts

Limitations. We present the limitations of our method in Fig. 13. These limitations can be mitigated by utilizing a heavier DiT-based video generation model, or a higher-quality motion extractor for motion-augmented rotary embedding generation.

A limitation of our proposed Fréchet Trajectory Distance is its reliance on CoTracker3 [24], which has difficulty handling zoom in or zoom out camera motions, especially in cases where the tracking points remain stationary. This limits the accuracy of the extracted trajectories in such scenarios.

Broader Impacts. The ability to generate realistic motion in videos can greatly benefit fields such as animation, virtual production, education, and accessibility. However, it also introduces risks, particularly in the creation of deepfakes and other forms of synthetic media that may be used

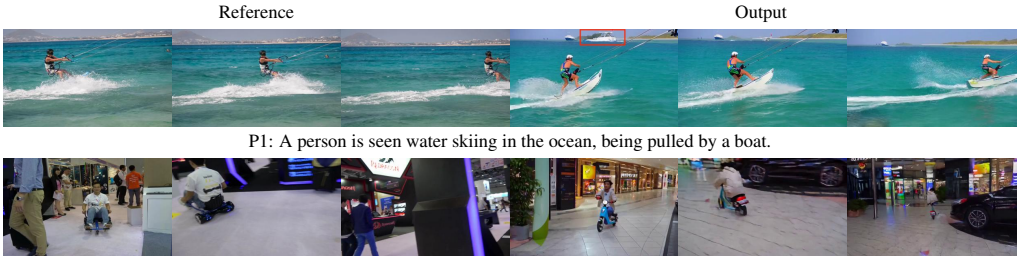


Figure 12: Limitations of our method. In the first video, a boat (highlighted with a red rectangle) intermittently appears and disappears, likely due to limitations in the backbone network. In the second video, the final frame shows a distorted human figure, caused by the absence of the person in the corresponding frame of the source video. This may highlight a limitation of the optical flow extractor used to modify our rotary embeddings, in handling occluded or missing subjects.

to deceive. These concerns highlight the importance of responsible use, and supporting research into detection and verification methods to help mitigate potential misuse while enabling positive applications.

A.6 More results on Challenging Videos

To further demonstrate the robustness of our method under challenging conditions, we also evaluate on a randomly sampled subsample from RealCamVid [49], each containing camera motion and multiple objects. Table 6 reveals that our method is also robust on more complex datasets than DAVIS.

Table 6: Comparison of motion transfer methods on RealCamVid [49].

Method	MF ↑	CLIP ↑	FTD ↓
GWTF	0.5796	0.2087	0.2072
ConMo	0.5973	0.2014	0.2800
SMM	0.5717	0.2118	0.2827
DiTFlow (RoPE)	0.5686	0.2118	0.2991
DiTFlow (latent)	0.5693	0.2099	0.2962
MOFT	0.5778	0.2095	0.2873
Ours	*0.7019*	0.2141	*0.1697*

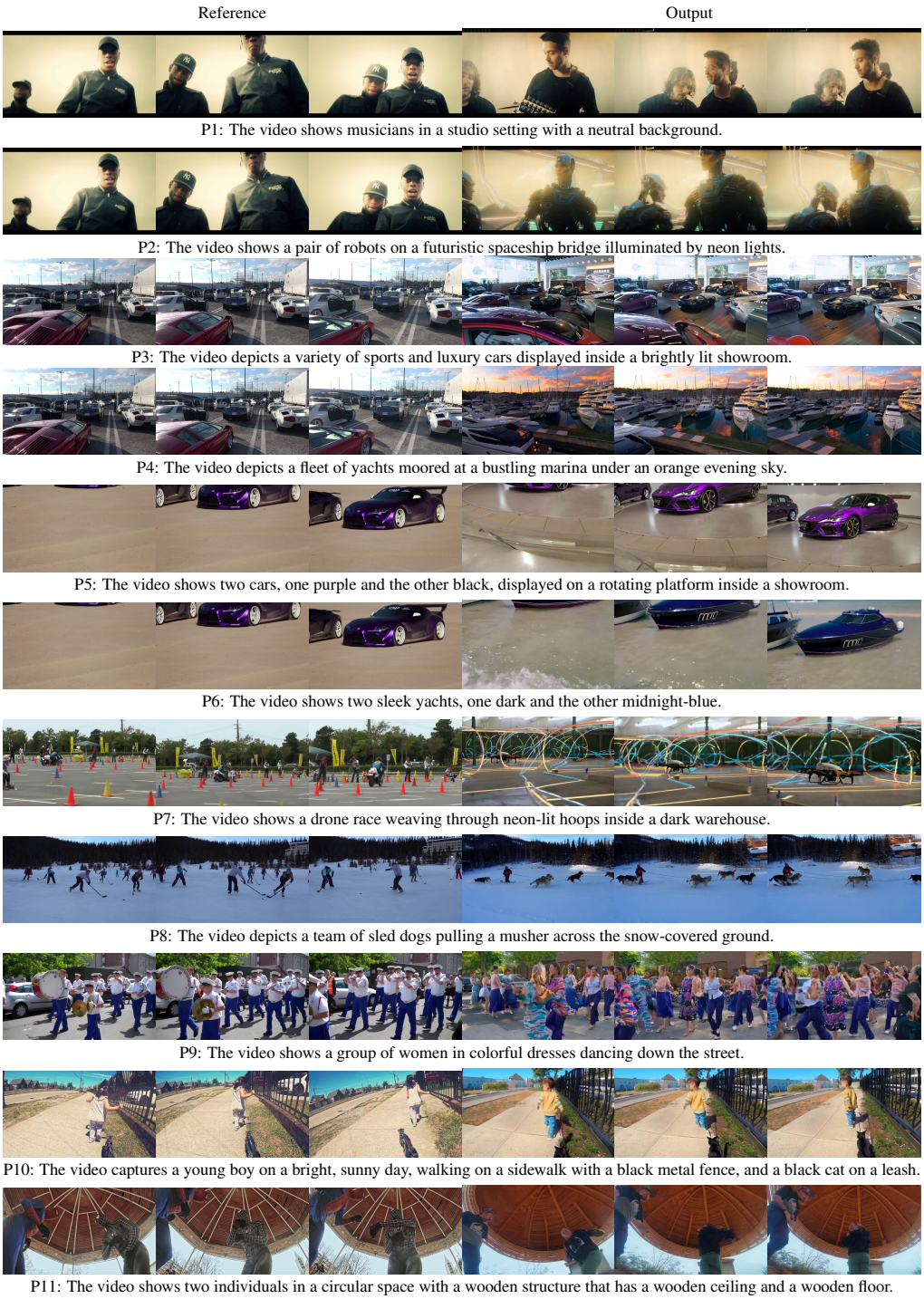


Figure 13: Our method effectively transfers camera motion and motion from multiple subjects accurately.