Eulerian Motion Guidance: Robust Image Animation via Bidirectional Geometric Consistency

Thong Nguyen*
Khoi M. Le
Cong-Duy Nguyen
Luu Anh Tuan
See-Kiong Ng
Chunyan Miao
thong.nguyen@u.nus.edu
National University of Singapore
Singapore
Centre for AI Research, VinUniversity
Vietnam
Nanyang Technological University
Singapore

Abstract

Recent advancements in image animation have utilized diffusion models to breathe life into static images. However, existing controllable frameworks typically rely on Lagrangian motion guidance, where optical flow is estimated relative to the initial frame. This paper revisits the same optical-flow primitive through a more local supervision design: we use adjacent-frame Eulerian motion fields to guide generation, where the motion signal always describes a short temporal hop. This shift enables parallelized training and provides bounded-error supervision throughout the generation process. To mitigate the drift artifacts common in adjacent frame generation, we introduce a Bidirectional Geometric Consistency mechanism, which computes a forward-backward cycle check to mathematically identify and mask occluded regions, preventing the model from learning incorrect warping objectives. Extensive experiments demonstrate that our approach accelerates training, preserves temporal coherence, and reduces dynamic artifacts compared to reference-based baselines.

CCS Concepts

• **Computing methodologies** → **Motion capture**; **Procedural animation**; • **Theory of computation** → **Theory and algorithms for application domains**.

Keywords

Image Animation, Motion Guidance, Geometric Consistency

1 Introduction

The pursuit of "*bringing images to life*" has long been a fascination in computer vision, evolving from early Generative Adversarial Network (GAN)-based approaches [26] to recent controllable diffusion models [12, 40]. While current Image-to-Video (I2V) frameworks such as Stable Video Diffusion [2] and Lumiere [1] can generate plausible short clips, ensuring robust long-term temporal consistency under complex motion remains a formidable challenge. The core difficulty lies in maintaining the identity and structure of the initial image while adhering to complex motion dynamics over time.

Existing methods [17, 22, 29] address this by integrating explicit motion control into frozen diffusion models. However, these approaches predominantly adopt a Lagrangian motion formulation, where optical flow is computed relative to the initial reference frame [8]. While conceptually simple, we identify two inherent structural vulnerabilities in this reference-anchored paradigm that make it highly susceptible to error accumulation. First, as the timestep increases, the displacement magnitude grows, often violating the brightness constancy assumptions inherent in optical flow estimation [25] and leading to increasingly noisy supervisory signals. Second, supervisory sparsity: as objects move or rotate, the valid correspondence area between the initial frame and the current frame decays exponentially due to occlusions, leaving the model hallucinating in unseen regions without guidance.

To mitigate these limitations, we propose a shift from Lagrangian-based to Eulerian-based Motion Guidance. Instead of tracking pixels relative to the start, we model motion as a dense field of transitions between consecutive frames. We provide a theoretical framework demonstrating how this formulation keeps optical flow estimates within the reliable, short-range regime of estimators, effectively bounding the per-step supervisory error. Moreover, this adjacent formulation allows for a parallelized batched flow computation strategy that minimizes sequential training overhead.

However, while Eulerian motion guidance provides stable local gradients, enforcing consistency only between neighboring frames can still lead to stochastic drift in regions that are newly revealed

*Corresponding Author

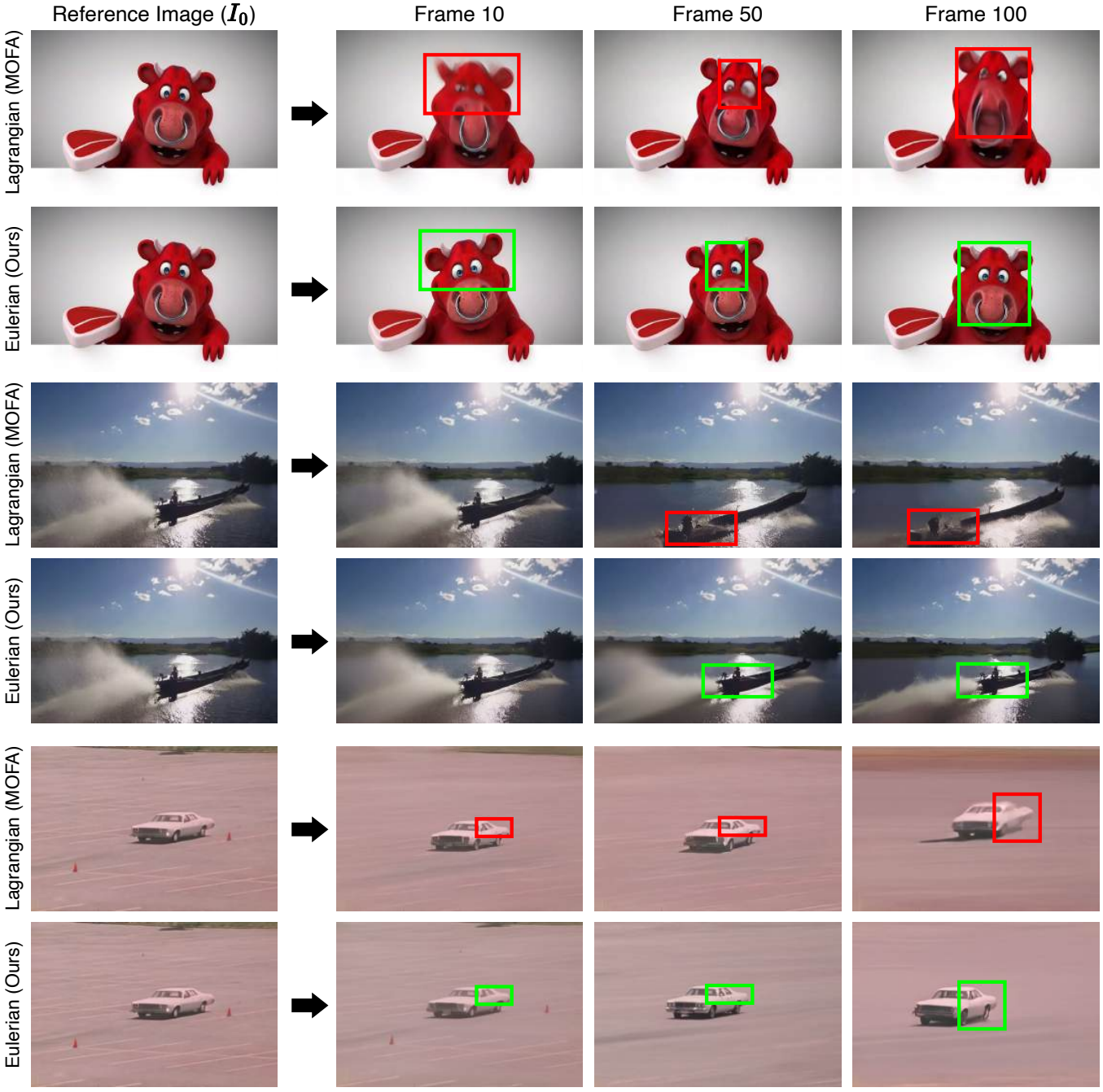

Figure 1: Long-horizon qualitative comparison. We show the reference image and generated frames at $t = \{10, 50, 100\}$. The Lagrangian baseline MOFA [22] exhibits increasing texture drift or blur as temporal distance grows, whereas our Eulerian motion framework better preserves identity and geometry over time.

over time. In these dis-occluded areas, there is no valid correspondence to the previous frame; forcing the model to guess the appearance often results in shimmering or ghosting artifacts. To address this inherent limitation of adjacent-frame tracking, we introduce a novel Bidirectional Geometric Consistency mechanism. By computing flow fields in both forward and backward directions, we

enforce a geometric cycle-consistency check, allowing us to mathematically derive an occlusion mask that dynamically filters out unreliable gradients in dis-occluded regions, ensuring that the model is supervised only by geometrically valid correspondences.

In summary, our contributions are three-fold:

- We formally analyze the error-accumulation vulnerabilities of Lagrangian motion guidance in video generation and propose an Eulerian alternative designed to bound per-step supervisory error.
- We introduce a Bidirectional Geometric Consistency mechanism that leverages forward-backward cycle checks to robustly handle occlusions and prevent artifact generation in dynamic scenes.
- We implement batched flow computation that integrates seamlessly with pre-trained diffusion backbones, achieving state-of-the-art performance in temporal consistency and visual fidelity.

2 Related Work

2.1 Controllable Video Generation.

Recent large-scale video generation models, such as Veo3 [30] and Wan2.1 [27], have demonstrated remarkable fidelity in text-to-video synthesis. However, fine-grained control remains a challenge. Existing controllable frameworks typically inject guidance via additional control layers or adapters. For trajectory-based control, early works like DragNUWA [35] and MotionCtrl [29] utilize sparse tracking points to guide generation. More recently, ATI [28] propose a unified trajectory instruction mechanism for camera and object motion, while PoseTraj [13] introduce pose-aware guidance to handle object rotations better. Despite these advancements, most prior methods rely on Lagrangian particle tracking-anchoring motion to the initial frame, which we argue leads to error accumulation in long-horizon generation. Go-with-the-flow [3] recently integrates optical flow into diffusion noise for motion control. However, unlike our explicit Eulerian flux, they operate via real-time noise warping, which can be sensitive to noise distribution shifts.

2.2 Portrait and Keypoint-Based Animation

Animating portraits from static images requires precise identity preservation. Foundational works in consistent animation, such as MagicAnimate [32], SadTalker [39], Cinemo [19], and Hallo3 [5], tackle consistent and controllable image animation using dense guidance and spatial-temporal attention mechanisms. Specialized architectures like EchoMimic [4] and FactorPortrait [24] achieve state-of-the-art disentanglement using implicit control. While these methods achieve high fidelity through domain-specific priors, our work demonstrates that a general-purpose Eulerian flow field can achieve competitive structural consistency across broader domains.

2.3 Geometric Consistency and Occlusion Handling

Maintaining temporal consistency in generative video, particularly under occlusion, is a long-standing problem. Traditional approaches in video magnification, such as MagFormer [10], have explored combining Lagrangian and Eulerian perspectives. Furthermore, previous generative pipelines have leveraged occlusion maps, such as LFDM [**?**] for latent flow diffusion and FlowVid [**?**] for forward-backward consistency in video editing. Confidence-based warping and occlusion-aware self-supervision have also been extensively explored in Video Frame Interpolation (VFI) [**?**].

While these methods utilize similar forward-backward checks, our Bidirectional Geometric Consistency (BGC) uniquely integrates this cycle-energy threshold directly into the diffusion training loop to gate dis-occluded gradients. By enforcing a forward-backward cycle check directly in the flow field, we achieve robust occlusion handling and texture preservation without the computational overhead of explicit 3D reconstruction seen in recent methods like Unboxed [37] or InsertAnywhere [14].

3 Preliminaries

Stable Video Diffusion. We employ *Stable Video Diffusion (SVD)* as our frozen generative backbone. SVD operates in a compressed latent space derived from a pre-trained VAE encoder, $z_0 = \mathcal{E}(V)$. The generation process reverses a noise perturbation chain via a 3D U-Net denoiser $S_\theta(z_t, t, c)$, which predicts the velocity field v_θ given a noisy latent z_t and a context embedding c, e.g., CLIP image embedding.

Controllable Image Animation. To inject motion control, we integrate a learnable motion adapter $\mathcal{M}$ into the frozen S_θ. The adapter transforms sparse user motion hints F^s into a dense flow field, which spatially aligns multi-scale reference features $f_0^{(k)}$ to the target timestep t via a differentiable warping operator $\mathcal{W}$. We optimize only the adapter parameters $\theta_\mathcal{M}$ using a standard reconstruction objective $\mathcal{L}_{\text{control}}$, keeping the backbone frozen.

4 Method

In this section, we delineate our Eulerian Motion Guidance (EMG) framework for image animation. We begin by establishing the theoretical intuition behind our approach through a formal analysis of expected error propagation in optical flow fields. Subsequently, we detail the Bidirectional Geometric Consistency regularizer, formulated as an energy minimization problem. Eventually, we demonstrate the parallelized motion computation for training efficiency and our inference pipeline for image animation.

4.1 From Lagrangian to Eulerian Motion Field

Most image animation frameworks guide motion in a Lagrangian manner: they estimate a reference-to-target displacement field from the reference frame at $t = 0$ to a future time t, and warp the reference latent accordingly. Concretely, given the reference image I_0 and its latent $z_0 = \mathcal{E}(I_0)$, they synthesize a motion-guided latent at time t via

$$\hat{z}_t^{\text{Lag}} = \mathcal{W}(z_0, \hat{\mathbf{u}}_{0\to t}), \tag{1}$$

where $\mathcal{W}(\cdot,\cdot)$ is a differentiable warping operator (e.g., bilinear sampling), and $\hat{\mathbf{u}}_{0\to t} : \Omega \to \mathbb{R}^2$ is the *estimated* displacement field defined on the reference coordinates.

Definition 1 (Lagrangian motion field). *Let $\Omega \subset \mathbb{R}^2$ denote the spatial domain of the reference frame at $t = 0$. The Lagrangian motion field is a reference-anchored displacement function $\mathbf{u}_{\text{Lag}} : \Omega \times \mathbb{R}^+ \to \mathbb{R}^2$ that maps each pixel coordinate $\mathbf{x} \in \Omega$ at $t = 0$ to its cumulative displacement at time t. The corresponding location at*

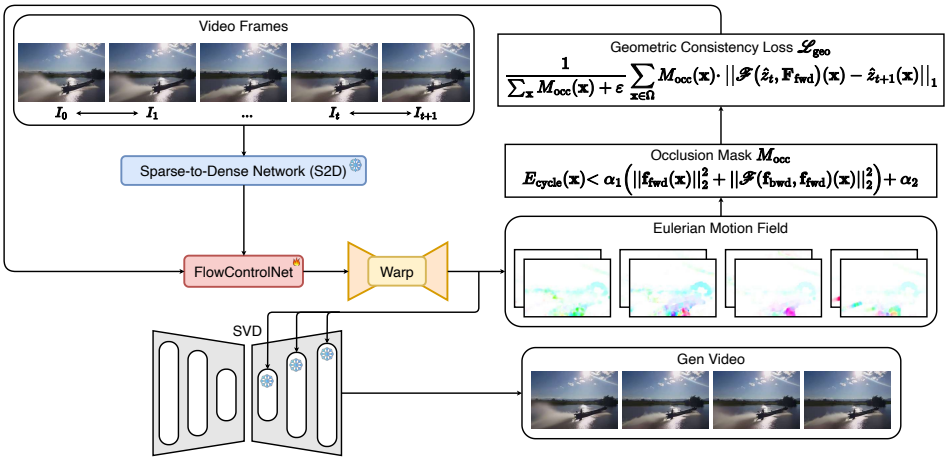

Figure 2: Eulerian Motion Guidance with Bidirectional Geometric Consistency. Given a reference image and sparse motion hints, a sparse-to-dense (S2D) module predicts an Eulerian motion field, which we then inject via Flow ControlNet by warping multi-scale reference features in a frozen SVD backbone. During training, forward and backward flows produce a cycle-based occlusion mask that gates geometric consistency loss, preventing supervision on dis-occluded regions.

time t is

$$\mathbf{x}_t = \mathbf{x} + \mathbf{u}_{\text{Lag}}(\mathbf{x}, t). \tag{2}$$

For discrete frames, the reference-to-t displacement used in warping is exactly $\mathbf{u}_{0\to t}(\mathbf{x}) \equiv \mathbf{u}_{\text{Lag}}(\mathbf{x}, t)$.

While conceptually simple, the reference-to-t formulation becomes increasingly vulnerable to error as t grows: (i) the displacement magnitude $\|\mathbf{u}_{0\to t}\|$ typically increases, making estimation and sampling more sensitive to error; and (ii) the set of reference pixels with valid correspondences, $\mathbb{X}_{\text{valid}}(t) \subseteq \mathbb{X}$, shrinks due to occlusion and out-of-bound mappings. Both effects amplify uncertainty in $\hat{\mathbf{u}}_{0\to t}$ and thus degrade the warped latent $\hat{z}_t^{\text{Lag}}$ over long horizons.

Theorem 1 (Expected Error Accumulation in Lagrangian Motion Fields). *Let $\mathbf{u}^*_{0\to t}$ be the ground-truth displacement from frame 0 to frame t. Under the standard assumption that the estimator $\hat{\mathbf{u}}_{0\to t} = \mathbf{u}^*_{0\to t} + \epsilon_t$ exhibits bounded kurtosis κ and an error variance that scales with the temporal baseline (due to growing deformation and decreasing correspondences), such that $\mathbb{E}[\epsilon_t] = \mathbf{0}$ and $\mathbb{E}[\|\epsilon_t\|^2] \geq \sigma^2 t$ for some $\sigma > 0$, the expected endpoint error is lower bounded by*

$$\mathbb{E}\left[\left\|\hat{\mathbf{u}}_{0\to t} - \mathbf{u}^*_{0\to t}\right\|\right] \;\geq\; \sigma\sqrt{t}. \tag{3}$$

We provide the proof of Theorem 1 in Appendix A. This growth implies that the supervisory target $\hat{z}_t^{\text{Lag}}$ becomes progressively less correlated with the true geometric structure at large t, forcing the denoising network to learn under increasingly high-variance guidance. To mitigate this instability, we switch from a Lagrangian motion field to an Eulerian one for image animation. Instead of predicting a single long-horizon displacement $\hat{\mathbf{u}}_{0\to t}$, we model motion as an instantaneous flux that transports the current state to the next state, and compose local transports over time.

Definition 2 (Eulerian Motion Field). *Let $\mathbb{X} \subset \mathbb{R}^2$ denote the spatial domain. The Eulerian motion field is a velocity function $\mathbf{u}_{\text{Eul}} : \mathbb{X} \times \mathbb{R}^+ \to \mathbb{R}^2$ defined at fixed coordinates, where $\mathbf{u}_{\text{Eul}}(\mathbf{x}, t)$ specifies the instantaneous motion (flux) observed at location $\mathbf{x} \in \Omega$ and time t. In continuous time, it induces pixel trajectories $\mathbf{x}(t)$ by the ODE*

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{u}_{\text{Eul}}(\mathbf{x}(t), t), \qquad \mathbf{x}(0) = \mathbf{x}_0. \tag{4}$$

For discrete frames, we represent this flux using adjacent-frame displacement (optical flow) $\mathbf{f}_{\tau\to\tau+1} : \mathbb{X} \to \mathbb{R}^2$, yielding the update

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{f}_{t\to t+1}(\mathbf{x}_t). \tag{5}$$

Let $\mathbb{X}_{\text{valid}}(t) \subseteq \mathbb{X}$ denote the set of pixels at time t that admit valid adjacent correspondences, *i.e.* not occluded and not mapped out of bounds under $\mathbf{f}_{t\to t+1}$. We now formalize how Eulerian supervision bounds the per-step gradient signal.

Theorem 2 (Uniform Bound on Eulerian Supervisory Error). *Let $\mathbf{f}^*_{t\to t+1}$ be the ground-truth adjacent-frame displacement, and let $\hat{\mathbf{f}}_{t\to t+1} = \mathbf{f}^*_{t\to t+1} + \epsilon_t$ be an estimator whose error satisfies $\mathbb{E}[\epsilon_t] = \mathbf{0}$ and*

$$\mathbb{E}\left[\frac{1}{|\mathbb{X}_{\text{valid}}(t)|}\int_{\mathbb{X}_{\text{valid}}(t)} \|\epsilon_t(\mathbf{x})\|^2 \, d\mathbf{x}\right] \;\leq\; \sigma^2, \qquad \forall t, \tag{6}$$

for some constant $\sigma > 0$ independent of t. Then the expected endpoint error of Eulerian supervision is uniformly bounded:

$$\mathbb{E}\left[\frac{1}{|\mathbb{X}_{\text{valid}}(t)|}\int_{\mathbb{X}_{\text{valid}}(t)} \left\|\hat{\mathbf{f}}_{t\to t+1}(\mathbf{x}) - \mathbf{f}^*_{t\to t+1}(\mathbf{x})\right\| \, d\mathbf{x}\right] \;\leq\; \sigma, \qquad \forall t. \tag{7}$$

We provide the proof of Theorem 2 in Appendix B. By minimizing the temporal interval to $\delta t = 1$, we ensure the displacement magnitude $\|\mathbf{f}_{t\to t+1}\|$ remains within the high-fidelity linear regime of the flow estimator, providing a stable, low-variance gradient signal throughout the entire generation window.

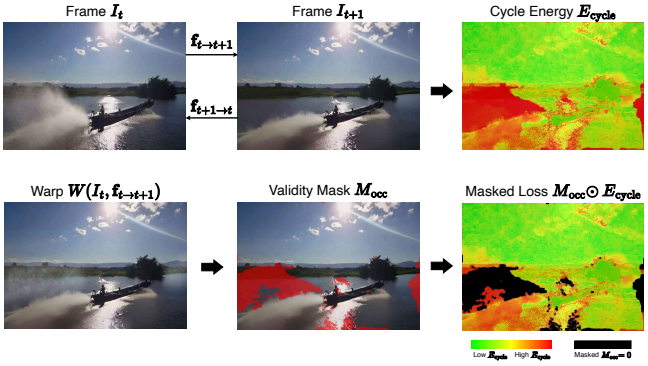

Figure 3: Occlusion masking from bidirectional cycle energy. For consecutive frames, we compute forward and backward flows, and evaluate cycle energy, for which high values indicate occlusion or flow failure. Thresholding yields a binary validity or occlusion mask, which is applied to the warp-based consistency loss so gradients are propagated only through geometrically verifiable correspondences.

4.2 Bidirectional Geometric Consistency

While Eulerian formulation bounds the local supervisory error, it introduces a risk of generative stochastic drift. Without a global anchor I_0, microscopic errors can accumulate autoregressively, leading to texture sliding in dis-occluded regions where the flow field $\mathbf{f}_{t\to t+1}$ is ill-defined. To regularize this, we propose the Bidirectional Geometric Consistency mechanism, formulated as a pixel-wise energy minimization problem.

Cycle Energy Formulation. We define the Cycle Consistency Energy $E_{\text{cycle}}(\mathbf{x})$ at coordinate $\mathbf{x} \in \mathbb{X}$ as the residual displacement magnitude after a forward-backward projection cycle:

$$E_{\text{cycle}}(\mathbf{x}) = \|\mathbf{f}_{t\to t+1}(\mathbf{x}) + \mathcal{F}\left(\mathbf{f}_{t+1\to t}, \mathbf{f}_{t\to t+1}\right)(\mathbf{x})\|_2^2, \tag{8}$$

where $\mathcal{F}(\mathbf{f}_{\text{bwd}}, \mathbf{f}_{\text{fwd}})$ denotes the backward flow vector sampled at the target location mapped by the forward flow. Non-zero energy, *i.e.* $E_{cycle} \gg 0$, indicates a violation of brightness constancy, serving as a robust proxy for occlusion or estimation failure.

Dynamic Occlusion Masking. We derive a binary validity mask $M_{\text{occ}} \in \{0,1\}^{H\times W}$ by treating the cycle energy as a random variable. To account for relative error tolerance, we employ a relaxed thresholding function parameterized by the local flow magnitude:

$$M_{\text{occ}}(\mathbf{x}) = \mathbb{I}\Big[E_{\text{cycle}}(\mathbf{x}) < \alpha_1\Big(\|\mathbf{f}_{t\to t+1}(\mathbf{x})\|_2^2 + \|\mathcal{W}(\mathbf{f}_{t+1\to t}, \mathbf{f}_{t\to t+1})(\mathbf{x})\|_2^2\Big) + \alpha_2\Big], \tag{9}$$

where $\mathbb{I}[\cdot]$ is the indicator function, and α_1, α_2 are hyperparameters governing the sensitivity of the decision boundary. This mask effectively partitions the spatial domain into the geometrically verifiable $\mathbb{X}_{\text{valid}}$ and the occluded $\mathbb{X}_{\text{occ}}$.

Geometric Consistency Loss. The final objective function integrates this geometric prior. We define the Geometric Consistency Loss $\mathcal{L}_{\text{geo}}$ as a spatially-weighted reconstruction objective, ensuring the generated latent $\hat{z}_{t+1}$ adheres to the Eulerian flux only within valid regions:

$$\mathcal{L}_{\text{geo}} = \frac{1}{\sum_{\mathbf{x}} M_{\text{occ}}(\mathbf{x}) + \epsilon} \sum_{\mathbf{x}\in\mathbb{X}} M_{\text{occ}}(\mathbf{x}) \cdot \|\mathcal{W}(\hat{z}_t, \mathbf{f}_{t\to t+1})(\mathbf{x}) - \hat{z}_{t+1}(\mathbf{x})\|_1. \tag{10}$$

By gating gradients with M_{occ}, we effectively sever the backpropagation graph in occluded regions, preventing the model from optimizing against hallucinating gradients. The total training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}$, where $\mathcal{L}_{\text{diff}}$ is the standard diffusion loss defined in Section 3.

4.3 Training Efficiency via Parallelized Motion Computation

While Eulerian formulation inherently implies a sequential dependency, strictly adhering to this autoregressive structure during training would incur a prohibitive linear computational cost $O(T)$. Because the groundtruth video $\mathcal{V}$ is fully observable during training, we exploit this non-causal availability to circumvent the sequential bottleneck.

We devise a parallelized motion computation strategy that reconfigures the input latent tensor $\mathcal{V} \in \mathbb{R}^{B\times T\times C\times H\times W}$ into a batched collection of temporal dyads. Particularly, we construct the sets of forward and backward pairs as $\mathcal{P}_{\text{fwd}} = \{(I_t, I_{t+1})\}_{t=0}^{T-2}$ and $\mathcal{P}_{\text{bwd}} = \{(I_{t+1}, I_t)\}_{t=0}^{T-2}$, respectively. By treating the temporal dimension as an extension of the batch dimension, we employ the shared, frozen flow estimator $\mathbb{X}(\cdot;\theta_{\mathbb{X}})$ to regress the bidirectional flow fields simultaneously:

$$\mathbf{F}_{\text{fwd}} = \mathbb{X}(\mathcal{P}_{\text{fwd}}) \in \mathbb{R}^{B(T-1)\times 2\times H\times W}, \tag{11}$$

$$\mathbf{F}_{\text{bwd}} = \mathbb{X}(\mathcal{P}_{\text{bwd}}) \in \mathbb{R}^{B(T-1)\times 2\times H\times W} \tag{12}$$

This operation effectively flattens the sequential bottleneck of motion computation. While the total computational work remains $O(T)$, this strategy enables $O(1)$ wall-clock training time relative to the sequence length, bounded only by the parallel hardware capacity. The resulting factor $\mathbf{F}_{\text{fwd}}$ is then fed into the Motion Adapter $\mathcal{M}$ to condition the denoising U-Net, while the bidirectional pair $(\mathbf{F}_{\text{fwd}}, \mathbf{F}_{\text{bwd}})$ facilitates geometric consistency.

4.4 Inference

During inference, we model the generation process as a conditional autoregressive sampling chain.

Sparse-to-Dense Warping. Given an initial frame I_0 and a set of sparse user trajectories $\mathcal{T} = \{(\mathbf{p}_i, \mathbf{v}_i)\}_{i=1}^N$, we first interpolate $\mathcal{T}$ to obtain a sparse hint map F_{sparse}. Motion adapter $\mathcal{M}$ acts as a generative prior to warp these sparse hints into Eulerian field $\hat{\mathbf{F}}_{\text{fwd}}$.

Autoregressive Generative Chain. The video sequence $V_{1:T}$ is generated via a Markov chain factorization:

$$p(V_{1:T}|I_0, \mathcal{C}) = \prod_{t=1}^{T} p(I_t|I_{t-1}, \hat{\mathbf{f}}_{t-1\to t}) \tag{13}$$

At each step t, the conditional probability $p(I_t|\cdot)$ is sampled via the reverse diffusion process. Specifically, the latent z_{t-1} is warped by the Eulerian flow $\hat{\mathbf{f}}_{t-1\to t}$ to serve as the structural condition for synthesizing z_t. Thanks to the bounded per-step supervisory error of our Eulerian formulation, coupled with bidirectional checks, this

autoregressive process prevents the structural degradation characteristic of Lagrangian warping, enabling the generation of consistent sequences with complex non-linear dynamics.

5 Experiments

5.1 Experimental Setup

Implementation Details. We adopt Stable Video Diffusion (SVD) [2] as a frozen generative backbone and train only the FlowControlNet on WebVid-10M using AdamW [18] with a learning rate of 2×10^{-5} at a resolution of 256×256. For our shared flow estimator $\mathcal{F}$, we utilize RAFT [25] pre-trained on FlyingThings3D, without further fine-tuning, operating at 256×256 resolution. The geometric consistency weights are empirically set to $\alpha_1 = 0.01$ and $\alpha_2 = 0.5$. All experiments are conducted on 4 NVIDIA H200 GPUs. At inference time, we perform image animation by generating sequences of $T = 100$ frames.

Evaluation Metrics. We evaluate our framework on both trajectory-based and keypoint-based image animation. For trajectory-based image animation, we randomly sample 1000 instances from the WebVid test set and report LPIPS [38], FID [11], and FVD [36]. Specifically, we compute FID and FVD by matching the distribution of our generated 100-frame sequences—decoded at a uniform 8 fps—directly against the ground-truth WebVid clips to ensure spatial and temporal distributions are strictly aligned. To assess temporal consistency, we compute the cosine similarity between CLIP embeddings [23] of consecutive generated frames. Because CLIP-Cons can inadvertently penalize intended large motions, we also report the Warping Error (E_{warp}) [15], which measures the pixel-level photometric error between consecutive frames after alignment via optical flow, thereby decoupling temporal consistency from motion magnitude. For keypoint-based image animation, we evaluate 40 generated videos, each of which contains 196 frames, then report Cumulative Probability Blur Detection (CPBD) [21] to measure sharpness, as well as identity preservation via ArcFace [6] between the source images and the generated frames. In addition to quantitative results, we also conduct user studies to compare perceptual quality against competing baselines.

5.2 Comparison with State-of-the-art Methods

In this section, we compare our results with the state-of-the-art methods tailored to each control setting. For all baselines, we follow the official implementations when available and keep the evaluation protocol identical across methods. Quantitative results are reported in Table 1 and Table 2. Qualitative results are provided in Figure 4 and Figure 5.

Trajectory-based Image Animation. We compare our method against recent trajectory-based video generation and image-to-video control methods, including 1) trajectory-guided diffusion fine-tuning approaches that directly inject sparse trajectories into the generation process, i.e. DragNUWA [35] and PoseTraj [13]; and 2) trajectory control methods built on top of the pretrained image-to-video backbones, i.e. MOFA [22], SG-I2V [20], I2VControl [7], AnyTraj [28], and ImageConductor [16]. Following prior works, we evaluate both object dragging and camera motion scenarios. For methods that require trajectory formats, e.g. bounding box trajectories or sparse point sets, we convert our user-specified trajectories into the required representation while preserving the same spatial path and temporal schedule. As shown in Table 1, we achieve substantially better perceptual scores, achieving the lowest LPIPS, FID, and FVD scores. Notably, we outperform the strongest baseline, i.e. ImageConductor [16], by a significant margin of 1.53 in FID and 3.02 in FVD, validating the efficacy of our parallelized Eulerian formulation in generating temporally coherent video distributions. Crucially, addressing recent concerns regarding the limitations of CLIP-Cons in measuring alignment under large motions, our framework achieves the lowest Warping Error ($E_{warp} = 1.84\times 10^{-3}$). This confirms that our Eulerian formulation and Bidirectional Geometric Consistency directly translate to superior pixel-level temporal stability, minimizing texture flickering even during rapid object translations.

In addition to quantitative comparison, qualitative comparison also reveals distinct advantages in long-horizon consistency, where Lagrangian-based methods typically struggle. In Figure 1, as temporal distance increases, MOFA exhibits severe textual drift and identity degradation. For instance, the structural identity of the red bear (row 1) and the boat's wake (row 2) collapses by frame 100 due to error accumulation in reference-anchored flow. In contrast, our Eulerian Motion Guidance maintains sharp structural details and clear object boundaries throughout the sequence. Moreover, in the jet-ski example (top row) of Figure 4, ImageConductor fails to maintain the rider's geometry, leading to the rider vanishing into the background by frame 50. Our method preserves the rider's silhouette against the fast-moving water. Furthermore, in the mixing bowl scene (bottom row), ImageConductor suffers from hallucination artifacts, generating unrealistic objects in the batter due to occlusion ambiguities. By leveraging Bidirectional Geometric Consistency to mask unreliable regions, our model generates plausible fluid dynamics without such hallucinations.

Keypoint-based Image Animation. We compare our method with strong portrait or talking-head baselines that specialize in keypoint-based image animation, including classical audio-to-motion pipelines and recent diffusion-based portrait animators, i.e. SadTalker [39], MagicAnimate [32], Cinemo [19], StyleHeat [34], Hallo [31], Hallo3 [5], FaceShot [9], and recent video-driven portrait methods, including Stable Video-Driven Portraits (SVDP) [33] and FactorPortrait [24]. Table 2 demonstrates that our method outperforms specialized portrait animators. We achieve the highest CPBD score, indicating superior frame sharpness, and the highest ArcFace score, reflecting better identity preservation than recent methods like Hallo3 [5], Cinemo [19], and FactorPortrait [24].

Furthermore, qualitative comparison in Figure 5 with FactorPortrait reveals our method's ability to retain fine facial details. In the sunglasses' subject (left), the baseline struggles with the dis-occluded mouth region, resulting in blurred teeth and undefined lip boundaries (highlighted in red boxes). Our method utilizes the Eulerian flux to propagate texture information from adjacent frames, yielding sharp, high-fidelity mouth movements (highlighted in green boxes). In the angry expression subject (right), our method maintains geometric stability under large deformations.

Table 1: Trajectory-based image animation on WebVid test set (1000 samples). Lower is better for LPIPS/FID/FVD/E_{warp}; higher is better for CLIP-Cons. Pref. is the user-study win-rate (%) of ours vs each baseline (95% CI in brackets).

Method	Backbone	LPIPS ↓	FID ↓	FVD ↓	CLIP-Cons ↑	E_{warp} ($\times 10^{-3}$) ↓	Pref. ↑
DragNUWA [35]	SVD	0.2705	19.66	91.38	0.9302	4.12	89.2 [85.1, 93.3]
PoseTraj [13]	SVD	0.1704	12.22	77.69	0.9562	2.58	63.5 [58.2, 68.8]
MOFA [22]	SVD	0.2274	16.82	86.76	0.9390	3.45	78.4 [73.5, 83.3]
SG-I2V [20]	SVD	0.2490	18.24	89.07	0.9346	3.81	82.1 [77.5, 86.7]
I2VControl [7]	MagicVideo-V2	0.2132	14.52	82.23	0.9476	3.10	71.3 [66.1, 76.5]
AnyTraj [28]	DiT	0.1989	13.75	80.71	0.9505	2.76	68.7 [63.2, 74.2]
ImageConductor [16]	AnimateDiff	0.1847	12.98	79.20	0.9534	2.51	65.2 [59.8, 70.6]
Ours	SVD	**0.1562**	**11.45**	**76.18**	**0.9591**	**1.84**	**92.3** [87.1, 97.5]

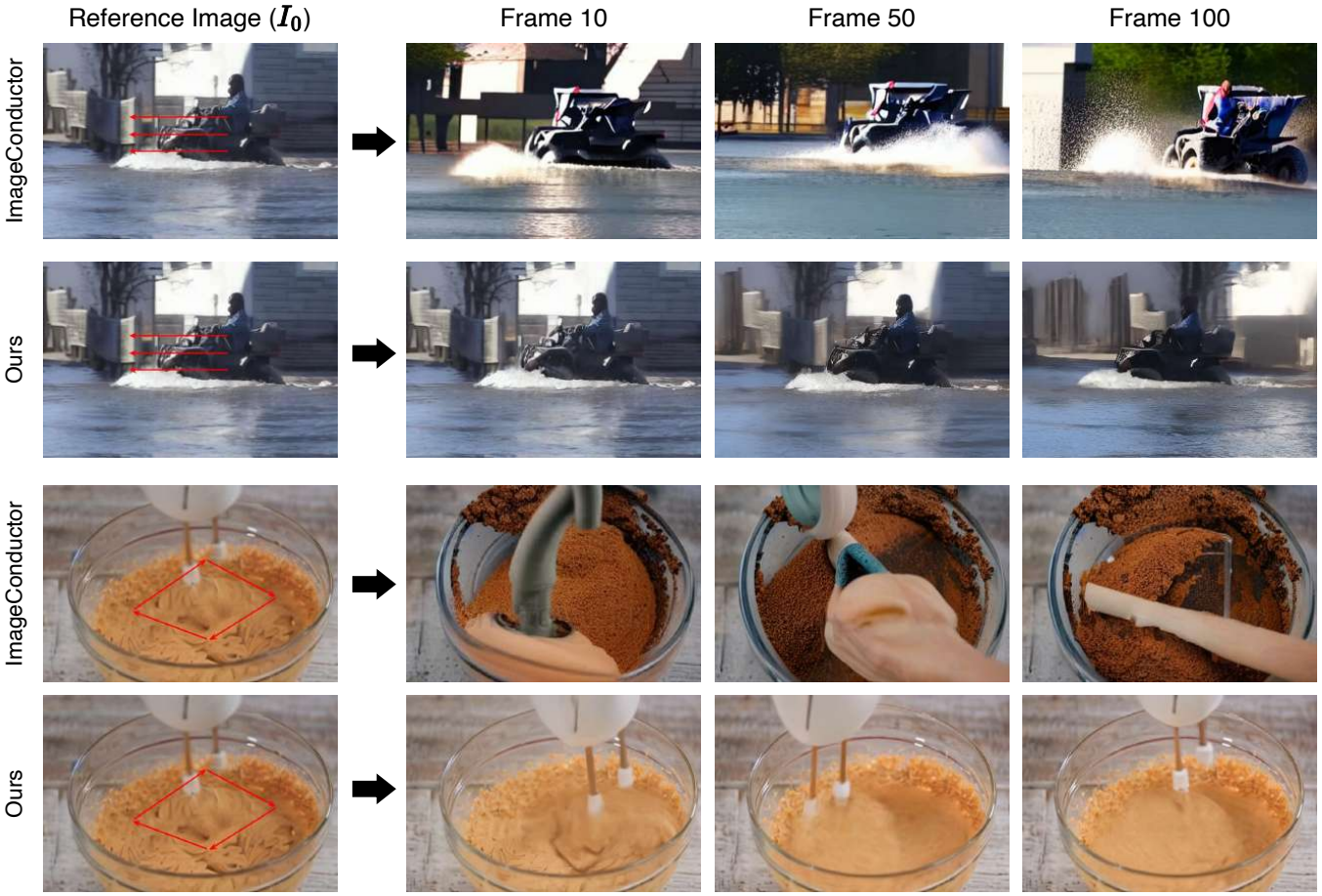

Figure 4: Qualitative comparison with ImageConductor [16] on long-horizon generation ($T = 100$). Red arrows in I_0 indicate the user-defined motion trajectories. Top: ImageConductor suffers from severe structural degradation and background drift, whereas our method preserves the rider's geometry. Bottom: ImageConductor hallucinates the mixing objects, e.g. hand or sponge, due to lack of valid correspondence. In contrast, our framework maintains texture fidelity throughout the complex circular motion.

While the baseline introduces temporal flicker and boundary artifacts around the jawline during intense expressions, our Bidirectional Geometric Consistency effectively filters inconsistent warping gradients, ensuring subject's identity remains stable during facial contortions.

User Studies. In addition to objective metrics, we conduct a user study to evaluate perceived visual quality, temporal stability, and control faithfulness. Participants are shown randomized pairwise comparisons between our method and each baseline under the same

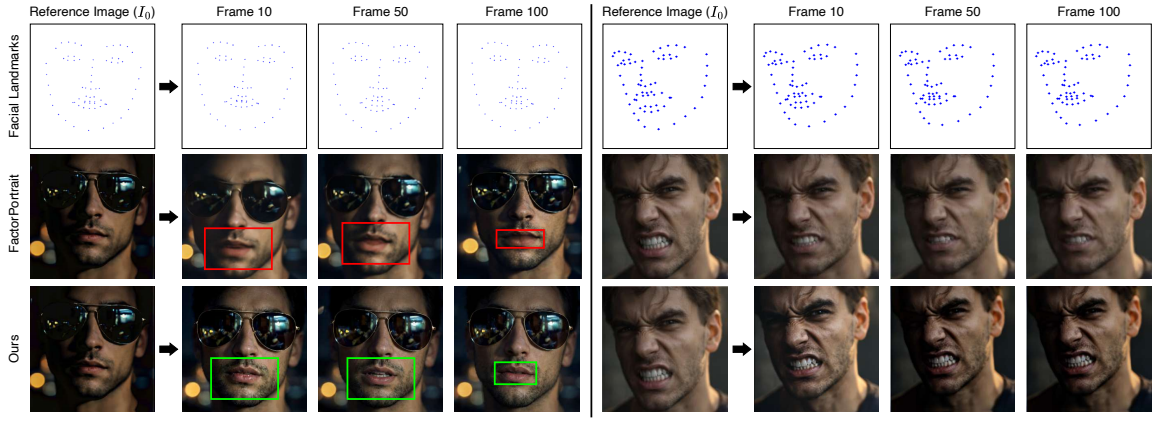

Figure 5: Qualitative comparison on keypoint-based animation. We compare our Eulerian Motion Guidance against the state-of-the-art baseline FactorPortrait [24]. Left: Our method preserves high-frequency details (e.g., mouth region) where the baseline suffers from texture blurring (highlighted). This validates our claim that Eulerian flux prevents the structural degradation typical of reference-anchored warping. Right: Our method maintains identity and geometric stability even under intense expression changes.

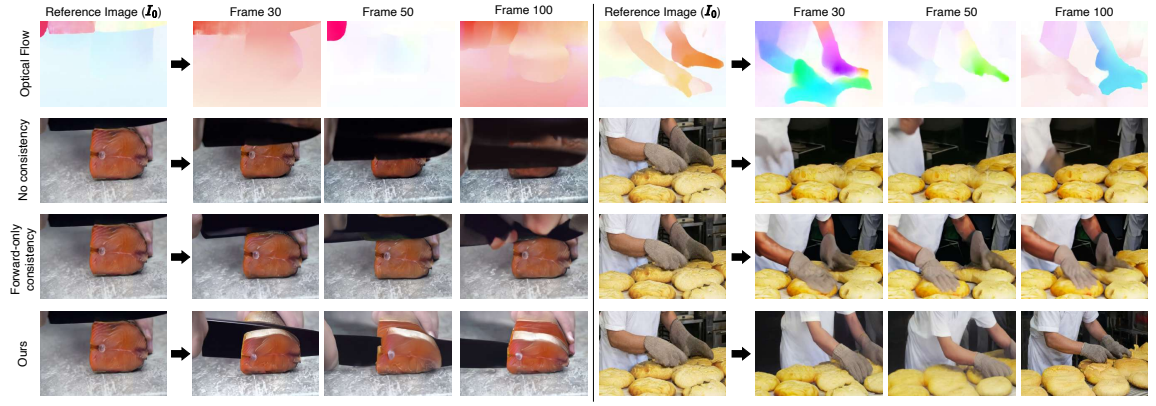

Figure 6: Qualitative ablation of geometric consistency. We compare training with (a) no consistency enforcement, (b) forward-only masking, and (c) our bidirectional cycle-based masking (BGC), under the same reference image and control signal, shown at frames t=30, 50, 100. Without reliable masking, dis-occluded regions produce ghosting artifacts where background textures adhere to moving foreground objects. Forward-only masking partially mitigates drift but fails under complex occlusions. BGC identifies unreliable correspondences via a forward–backward cycle check and suppresses incorrect warping supervision, preserving structure and appearance over long horizons.

input and control signals. We ask them to select the preferred output, and report win rates and confidence intervals in Table 1 and 2. We can observe that the volunteers prefer the proposed framework obtains the highest preferred rates across all approaches.

5.3 Ablation Study

To validate the effectiveness of our proposed contributions, we conduct ablation studies on the WebVid test set. We analyze three core aspects: training efficiency, motion guidance formulation, and geometric consistency.

Training Efficiency. A key advantage of our Eulerian formulation is the ability to parallelize motion estimation (Sec. 4.3). We compare the training iteration time of our method against a standard sequential implementation where flow is estimated autoregressively. Figure 7 shows that the sequential baseline scales linearly $\mathcal{O}(T)$, creating a computational bottleneck for long sequences. In contrast, our parallelized strategy maintains nearly constant time complexity $\mathcal{O}(1)$ (relative to temporal hardware capacity). For a sequence length of $T = 24$, our method reduces iteration time from 315ms to 115ms, achieving a 2.7× speedup.

Effectiveness of Eulerian Motion Guidance. We train a Lagrangian variant that predicts displacements relative to the initial frame I_0 while keeping the SVD backbone and all other settings identical. As shown in Table 3, the Lagrangian baseline degrades significantly

Table 2: Keypoint-based portrait animation comparison. Lower is better for E_{warp}; Higher is better for CPBD/ArcFace/CLIP-Cons.

Method	CPBD ↑	ArcFace ↑	CLIP-Cons ↑	E_{warp} ($\times 10^{-3}$) ↓	Pref. ↑
SadTalker [39]	0.3218	0.9188	0.9156	3.88	88.5 [84.1, 92.9]
MagicAnimate [32]	0.3852	0.9241	0.9288	2.95	83.7 [78.5, 88.4]
Cinemo [19]	0.3986	0.9315	0.9342	2.61	77.1 [71.8, 82.3]
MOFA [22]	0.4075	0.9293	0.9390	2.45	79.2 [74.3, 84.1]
Hallo [31]	0.3912	0.9331	0.9364	2.38	75.4 [70.1, 80.7]
Hallo3 [5]	0.4045	0.9389	0.9409	2.21	72.8 [67.2, 78.4]
FaceShot [9]	0.4178	0.9439	0.9455	2.15	68.1 [62.4, 73.8]
SVDP [33]	0.4310	0.9485	0.9486	2.08	64.3 [58.5, 70.1]
FactorPortrait [24]	0.4343	0.9491	0.9459	2.10	61.7 [55.9, 67.5]
Ours	**0.4576**	**0.9558**	**0.9591**	**1.52**	**94.4** [89.0, 99.7]

Table 3: Ablation Study. We evaluate the contribution of Eulerian Motion Guidance and Bidirectional Geometric Consistency (BGC) on the WebVid test set. Lagrangian denotes a baseline trained with reference-to-target flow. Forward-only uses a naive forward flow magnitude mask. Pref. reports the *variant win-rate* (%) against our full model (Eulerian + BGC) in a pairwise user study (lower is better).

Model Variant	Consistency Strategy	LPIPS ↓	FVD ↓	CLIP-Cons ↑	Pref.
Lagrangian Baseline	None	0.224	85.12	0.935	12.5%
Eulerian	None	0.178	79.45	0.951	28.3%
Eulerian	Forward-only Mask	0.165	77.80	0.955	41.2%
Eulerian (Ours)	**Bidirectional (BGC)**	**0.156**	**76.18**	**0.959**	–

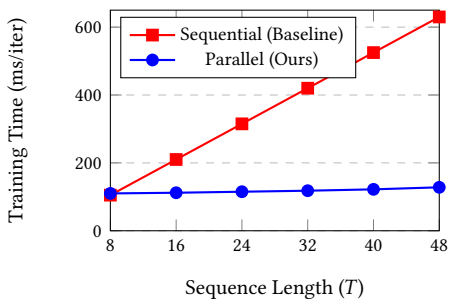

Figure 7: Training Efficiency Analysis. comparison of per-iteration training time between standard sequential flow estimation (Red) and our parallelized strategy (Blue). Our method achieves $\mathcal{O}(1)$ complexity relative to sequence length, yielding a ~3× speedup at $T = 24$.

(LPIPS rises to 0.224, FVD to 85.12, CLIP-Cons drops to 0.935), confirming that reference-anchored flow suffers from error accumulation and vanishing valid correspondences over long horizons. Our Eulerian formulation bounds per-step error (Theorem 1 and 2), preserving sharp textures and structural fidelity throughout 100-frame sequences.

Impact of Bidirectional Geometric Consistency. The Eulerian formulation is prone to drift in dis-occluded regions. We compare three strategies on the Eulerian baseline: (1) no consistency check (full-pixel supervision), (2) forward-only magnitude masking, and (3) our BGC using forward-backward cycle energy E_{cycle}. Removing consistency increases LPIPS to 0.178. The forward-only heuristic helps modestly but fails under rotation and complex occlusions. BGC achieves the best scores (LPIPS 0.156, FVD 76.18, CLIP-Cons 0.959) by gating unreliable gradients. Figure 6 shows that BGC eliminates ghosting and boundary artifacts, cleanly separating foreground from newly revealed background. More comprehensive analysis can be found in Appendix D.

6 Conclusion

In this paper, we identify a fundamental theoretical limitation in existing image animation frameworks, i.e. the unbounded error growth inherent in reference-anchored Lagrangian motion guidance. To address this, we propose Eulerian Motion Guidance, a novel paradigm that formulates animation as a sequence of adjacent, bounded-error flow transitions. We further introduce the Bidirectional Geometric Consistency mechanism, which utilizes a forward-backward cycle energy check to mathematically detect and mask unreliable gradients in dis-occluded regions. Extensive experiments on trajectory-based and keypoint-based tasks demonstrate that our approach significantly outperforms recent state-of-the-art methods. Our method not only generates sharper textures over long horizons but also eliminates the artifacts common in complex dynamic scenes.

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

[3] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. 2025. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13–23.

[4] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. 2025. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2403–2410.

[5] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. 2025. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 21086–21095.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[7] Wanquan Feng, Tianhao Qi, Jiawei Liu, Mingzhen Sun, Pengqi Tu, Tianxiang Ma, Fei Dai, Songtao Zhao, Siyu Zhou, and Qian He. 2025. I2vcontrol: Disentangled and unified video motion synthesis control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14051–14060.

[8] Craig G Fraser. 2005. Leonhard Euler, book on the calculus of variations (1744). In *Landmark Writings in Western Mathematics 1640-1940*. Elsevier, 168–180.

[9] Junyao Gao, Yanan Sun, Fei Shen, Xin Jiang, Zhening Xing, Kai Chen, and Cairong Zhao. 2025. Faceshot: Bring any character into life. *arXiv preprint arXiv:2503.00740* (2025).

[10] Sicheng Gao, Yutang Feng, Linlin Yang, Xuhui Liu, Zichen Zhu, David S Doermann, and Baochang Zhang. 2022. MagFormer: Hybrid Video Motion Magnification Transformer from Eulerian and Lagrangian Perspectives.. In *BMVC*. 444.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[12] Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8153–8163.

[13] Longbin Ji, Lei Zhong, Pengfei Wei, and Changjian Li. 2025. PoseTraj: Pose-Aware Trajectory Control in Video Diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 22776–22785.

[14] Hoiyeong Jin, Hyojin Jang, Jeongho Kim, Junha Hyung, Kinam Kim, Dongjin Kim, Huijin Choi, Hyeonji Kim, and Jaegul Choo. 2025. InsertAnywhere: Bridging 4D Scene Geometry and Diffusion Models for Realistic Video Object Insertion. *arXiv preprint arXiv:2512.17504* (2025).

[15] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*. 170–185.

[16] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. 2025. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5031–5038.

[17] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. 2024. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*. Springer, 56–74.

[18] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[19] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024. Cinemo: Consistent and controllable image animation with motion diffusion models. *arXiv preprint arXiv:2407.15642* (2024).

[20] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. 2024. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989* (2024).

[21] Niranjan D Narvekar and Lina J Karam. 2011. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Transactions on Image Processing* 20, 9 (2011), 2678–2683.

[22] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. 2024. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*. Springer, 111–128.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR.

[24] Jiapeng Tang, Kai Li, Chengxiang Yin, Liuhao Ge, Fei Jiang, Jiu Xu, Matthias Nießner, Christian Häne, Timur Bagautdinov, Egor Zakharov, et al. 2025. FactorPortrait: Controllable Portrait Animation via Disentangled Expression, Pose, and Viewpoint. *arXiv preprint arXiv:2512.11645* (2025).

[25] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*. Springer, 402–419.

[26] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. *Advances in neural information processing systems* 29 (2016).

[27] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).

[28] Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. 2025. ATI: Any Trajectory Instruction for Controllable Video Generation. *arXiv preprint arXiv:2505.22944* (2025).

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

[30] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. 2025. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328* (2025).

[31] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. 2024. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801* (2024).

[32] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1481–1490.

[33] Fei Yin, Vikram Voleti, Nikita Drobyshev, Maksim Lapin, Aaryaman Vasishta, Varun Jampani, et al. 2025. Stable Video-Driven Portraits. *arXiv preprint arXiv:2509.17476* (2025).

[34] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. 2022. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*. Springer, 85–101.

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

[36] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. 2022. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv preprint arXiv:2202.10571* (2022).

[37] Zhongrui Yu, Martina Megaro-Boldini, Robert W Sumner, and Abdelaziz Djelouah. 2025. Unboxed: Geometrically and Temporally Consistent Video Outpainting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7309–7319.

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

[39] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8652–8661.

[40] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22490–22499.

A Proof of Theorem 1

Theorem 1. (Expected Error Accumulation in Lagrangian Motion Fields). *Let $\mathbf{u}^*_{0\to t}$ be the ground-truth displacement from frame 0 to frame t. Under the standard assumption that the estimator $\hat{\mathbf{u}}_{0\to t} = \mathbf{u}^*_{0\to t} + \epsilon_t$ exhibits an error variance that scales with the temporal baseline (due to growing deformation and decreasing correspondences), such that $\mathbb{E}[\epsilon_t] = \mathbf{0}$ and $\mathbb{E}[\|\epsilon_t\|^2] \geq \sigma^2 t$ for some $\sigma > 0$, the expected endpoint error is lower bounded by*

$$\mathbb{E}\big[\big\|\hat{\mathbf{u}}_{0\to t} - \mathbf{u}^*_{0\to t}\big\|\big] \;\geq\; \sigma\sqrt{t}. \tag{14}$$

Proof. Let $Z = \|\epsilon_t\|$ denote the error magnitude. By our variance-scaling assumption, the second moment satisfies $\mathbb{E}[Z^2] \geq \sigma^2 t$. While a large second moment implies a large Mean Squared Error (MSE), it does not strictly imply a large expected L1 error ($\mathbb{E}[Z]$) without constraints on the distribution's heavy-tailedness. We utilize the Paley-Zygmund inequality to strictly lower bound the first moment.

Let $X = Z^2$. Using the Paley-Zygmund inequality with $\theta = 0.5$, we have:

$$\mathbb{P}(X > 0.5\mathbb{E}[X]) \geq \frac{(1-0.5)^2\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{15}$$

Using the bounded kurtosis assumption $\mathbb{E}[X^2] = \mathbb{E}[Z^4] \leq \kappa(\mathbb{E}[Z^2])^2 = \kappa\mathbb{E}[X]^2$, we obtain:

$$\mathbb{P}(Z^2 > 0.5\mathbb{E}[Z^2]) \geq \frac{0.25}{\kappa} = \frac{1}{4\kappa}. \tag{16}$$

This inequality states that with probability at least $\frac{1}{4\kappa}$, the squared error is significant. We can now lower bound the expected error:

$$\begin{aligned}
\mathbb{E}[Z] &\geq \mathbb{E}[Z \cdot \mathbb{I}(Z^2 > 0.5\mathbb{E}[Z^2])] \\
&\geq \sqrt{0.5\mathbb{E}[Z^2]} \cdot \mathbb{P}(Z^2 > 0.5\mathbb{E}[Z^2]) \\
&\geq \frac{1}{\sqrt{2}}\sigma\sqrt{t} \cdot \frac{1}{4\kappa} \\
&= \frac{1}{4\sqrt{2}\kappa}\sigma\sqrt{t}.
\end{aligned} \tag{17}$$

As $t \to \infty$, the term $\sigma\sqrt{t} \to \infty$. Thus, if the error variance scales with time, the expected endpoint error of the Lagrangian estimator diverges at a rate of $\mathbb{X}(\sqrt{t})$, explaining the empirical drift observed in long-horizon generation. □

B Proof of Theorem 2

Theorem 2. (Uniform Bound on Eulerian Supervisory Error). *Let $\mathbf{f}^*_{t\to t+1}$ be the ground-truth adjacent-frame displacement, and let $\hat{\mathbf{f}}_{t\to t+1} = \mathbf{f}^*_{t\to t+1} + \epsilon_t$ be an estimator whose error satisfies $\mathbb{E}[\epsilon_t] = \mathbf{0}$ and*

$$\mathbb{E}\left[\frac{1}{|\mathbb{X}_{\text{valid}}(t)|}\int_{\mathbb{X}_{\text{valid}}(t)} \|\epsilon_t(\mathbf{x})\|^2 \, d\mathbf{x}\right] \;\leq\; \sigma^2, \qquad \forall t, \tag{18}$$

for some constant $\sigma > 0$ independent of t. Then the expected endpoint error of Eulerian supervision is uniformly bounded:

$$\mathbb{E}\left[\frac{1}{|\mathbb{X}_{\text{valid}}(t)|}\int_{\mathbb{X}_{\text{valid}}(t)} \left\|\hat{\mathbf{f}}_{t\to t+1}(\mathbf{x}) - \mathbf{f}^*_{t\to t+1}(\mathbf{x})\right\| d\mathbf{x}\right] \;\leq\; \sigma, \qquad \forall t. \tag{19}$$

Proof. Define the pointwise flow error field

$$\varepsilon_t(x) := \hat{f}_{t\to t+1}(x) - f^*_{t\to t+1}(x), \quad x \in \mathbb{X}_{\text{valid}}(t).$$

Let

$$A_t := \langle\|\varepsilon_t(\cdot)\|\rangle_{\mathbb{X}_{\text{valid}}(t)} = \frac{1}{|\mathbb{X}_{\text{valid}}(t)|}\int_{\mathbb{X}_{\text{valid}}(t)} \|\varepsilon_t(x)\| \, dx.$$

By Cauchy–Schwarz on $\mathbb{X}_{\text{valid}}(t)$,

$$\left(\int_{\mathbb{X}_{\text{valid}}(t)} \|\varepsilon_t(x)\| \, dx\right)^2 \leq |\mathbb{X}_{\text{valid}}(t)| \int_{\mathbb{X}_{\text{valid}}(t)} \|\varepsilon_t(x)\|^2 \, dx.$$

Dividing both sides by $|\mathbb{X}_{\text{valid}}(t)|^2$ yields the deterministic inequality

$$A_t^2 \;\leq\; \langle\|\varepsilon_t(\cdot)\|^2\rangle_{\mathbb{X}_{\text{valid}}(t)}.$$

Taking expectation and using the assumption of Theorem 2 gives

$$\mathbb{E}[A_t^2] \leq \mathbb{E}\Big[\langle\|\varepsilon_t(\cdot)\|^2\rangle_{\mathbb{X}_{\text{valid}}(t)}\Big] \leq \sigma^2.$$

Finally, since $A_t \geq 0$, Jensen (equivalently $\mathbb{E}[A_t] \leq \sqrt{\mathbb{E}[A_t^2]}$) implies

$$\mathbb{E}[A_t] \;\leq\; \sqrt{\mathbb{E}[A_t^2]} \;\leq\; \sigma,$$

which is exactly the claimed uniform bound for all t. □

C More Visual Results

We provide extended illustrations to further examine the behavior of our method under diverse motion conditions in Figure 8 and 9.

D Sensitivity Analysis of Geometric Consistency

As introduced in Section 4.2, our Bidirectional Geometric Consistency (BGC) mask relies on a dynamic threshold defined by α_1 and α_2 to identify valid correspondences. To validate the robustness of these hyperparameters and understand their interaction with motion magnitude and texture complexity, we conduct a comprehensive sensitivity analysis.

Hyperparameter Robustness. The threshold balances a dynamic motion-dependent term (α_1) and a static noise floor (α_2). In Table 4, we vary $\alpha_1 \in \{0.005, 0.01, 0.05\}$ and $\alpha_2 \in \{0.1, 0.5, 1.0\}$ and report the resulting LPIPS and FVD on a 200-video subset of the WebVid test set. We observe that our framework is relatively stable across a moderate range of values. Setting α_2 too low (0.1) makes the mask overly aggressive, discarding valid gradients and slightly degrading FVD, whereas setting α_1 too high (0.05) makes the mask too permissive, allowing forward-backward drift to contaminate the training signal, which increases LPIPS. The empirical choice of $\alpha_1 = 0.01$ and $\alpha_2 = 0.5$ provides the optimal trade-off between structural preservation and training efficiency.

Interaction with Motion Magnitude. The dynamic formulation of the occlusion threshold is specifically designed to adapt to varying motion scales . In regions with extreme motion trajectories, the inherent interpolation variance of the flow estimator $\mathbb{X}$ increases, leading to a naturally higher cycle error even in non-occluded regions. By scaling the threshold with the flow magnitude via $\alpha_1\|\mathbf{f}\|_2^2$, the mask automatically relaxes its strictness for fast-moving objects, preventing the erroneous masking of high-velocity foreground elements. As demonstrated in our trajectory-based section 4.2, this

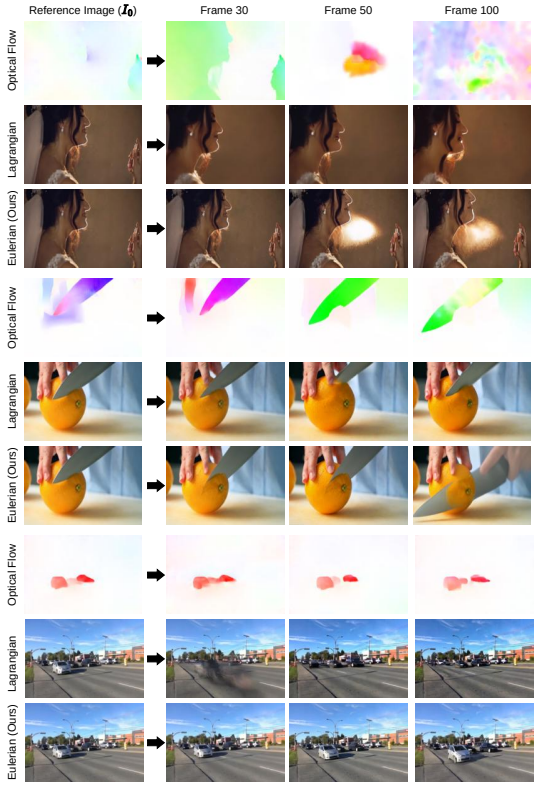

Figure 8: Robustness to Large Displacement. We compare Eulerian Motion Guidance against the Lagrangian baseline on open-domain video generation. Row 2 visualizes the optical flow magnitude, illustrating high-motion regions. In the Lagrangian baseline (Row 3), the structure of the subject degrades significantly by Frame 100 due to drift. In contrast, our Eulerian approach (Row 4) maintains sharp boundaries and coherent textures throughout the generation window.

Table 4: Sensitivity Analysis of BGC Thresholds. Evaluation of LPIPS (↓) and FVD (↓) under varying α_1 and α_2 values. The model is highly robust near our default settings ($\alpha_1 = 0.01, \alpha_2 = 0.5$).

α_1 (Dynamic)	α_2 (Static)	LPIPS ↓	FVD ↓
0.005	0.5	0.1581	77.05
0.010	**0.5**	**0.1562**	**76.18**
0.050	0.5	0.1634	78.42
0.010	0.1	0.1578	79.11
0.010	1.0	0.1610	77.85

prevents the "*vanishing subject*" artifacts commonly seen in fixed-threshold Eulerian approaches.

Interaction with Texture Complexity. The static threshold component, α_2, acts as an essential noise floor for texture variations. In regions with high-frequency textures, e.g., foliage, intricate clothing, sub-pixel flow estimation becomes inherently noisy. Conversely, in textureless regions, e.g., plain walls, clear skies, the aperture problem causes flow ambiguity, leading to small but non-zero forward-backward cycle errors. The parameter α_2 ensures that these minor, texture-induced flow variations are not incorrectly flagged as dis-occlusions. During our qualitative analysis, removing α_2 resulted

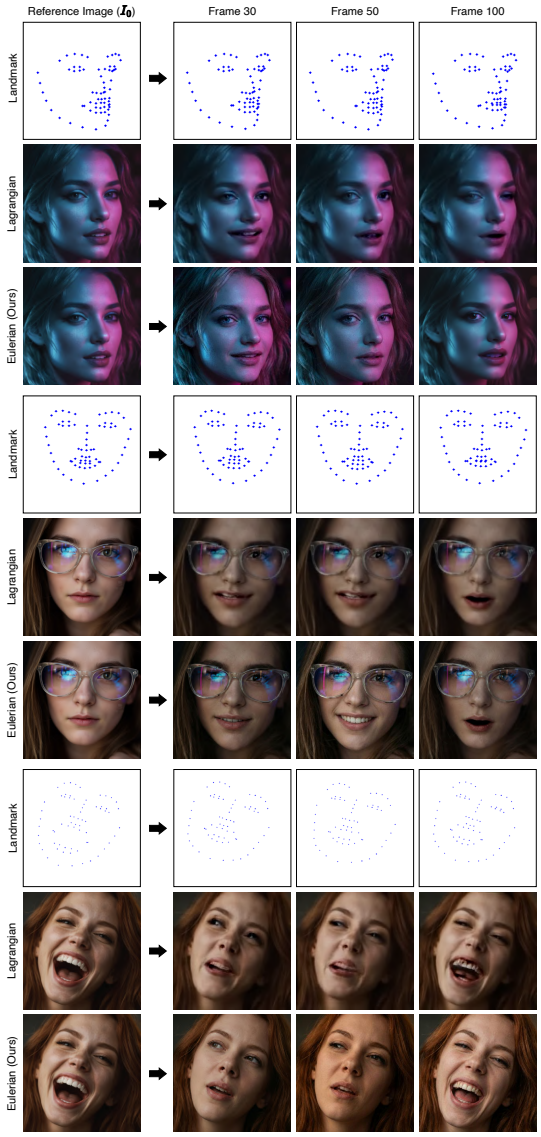

Figure 9: Extended Qualitative Evaluation on Landmark-Guided Portrait Animation. We compare our method against a standard Lagrangian baseline driven by the same sparse facial landmarks. (Top Group) Under complex lighting, the Lagrangian baseline loses skin texture detail and identity by Frame 100, whereas our method preserves the subject's likeness. (Middle Group) The Lagrangian approach struggles with rigid objects attached to the face; note the severe distortion of the eyeglasses in the later frames (Lagrangian) compared to the geometric stability maintained by our Eulerian method. (Bottom Group) In high-dynamic expressions (laughing), our method prevents the "blurring" of teeth and mouth interior often observed in reference warping.

in a highly fragmented gradient mask, leading to localized blurring in textured regions due to intermittent supervision.