AlignMorph: Tuning-Free Diffusion Image Morphing via Explicit Semantic Transport

Wuyi Liu[1], Xu Han[1], Yuren Chen[2], Yige Mao[3], and Xianzhi Li[1⋆]

[1] Huazhong University of Science and Technology
[2] Beijing Jiaotong University
[3] Beihang University

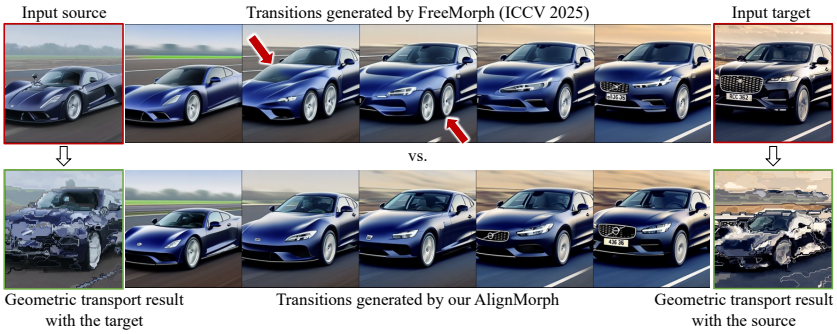

Fig. 1: Comparisons between our AlignMorph and the very recent FreeMorph [7] given two distinct inputs. See the two images marked by green boxes, our explicit semantic alignment achieves accurate geometry transport at the semantic level instead of traditional pixel-level matching, leading to smoother and more reasonable transitions.

Abstract. Image morphing aims to produce a smooth and semantically consistent transition between two input images. Existing diffusion-based morphing methods either require expensive per-pair optimization or rely on implicit spatial alignment, which easily fails under large layout discrepancies. To address these limitations, we propose **AlignMorph**, a novel tuning-free diffusion framework guided by the principle of *transport-then-denoise*. We explicitly decouple geometric alignment from generative denoising to avoid structural entanglement. Our framework consists of two core components. (1) Global Semantic Transport, which achieves diffusion-compatible semantic alignment via entropic optimal transport and reliability-aware latent warping; and (2) Coordinate-Aligned Generation, which uses a symmetric bi-phase attention handoff to maintain consistent spatial coordinates throughout denoising. Without any tuning, AlignMorph effectively eliminates ghosting and achieves superior structural coherence and temporal smoothness

⋆ Corresponding author

on morphing benchmarks. Code is available at `https://github.com/51xOne/Alignmorph`.

Keywords: Image Morphing · Diffusion Models · Semantic Correspondence

1 Introduction

Image morphing aims to generate a smooth visual transition between two distinct images through geometric warping and color blending, serving as a pivotal technique for animation and creative workflows [53]. Unlike standard text-to-video generation, image morphing is jointly constrained by the input source and target images. It requires not only smooth and continuous transitions between intermediate frames, but also semantic consistency with both input images.

To address this challenging task, generative diffusion models have been increasingly adopted as the dominant engines, building upon their extensive capabilities in high-quality image synthesis, editing, and video generation [38, 42]. Recent diffusion-based morphing methods can be broadly categorized into optimization-based and tuning-free approaches. Optimization-based methods such as DiffMorpher [55] and IMPUS [54] achieve strong fidelity by optimizing LoRA modules or latent embeddings for each image pair, but the per-instance tuning incurs substantial computational cost and limits practical use. To remove this bottleneck, tuning-free methods represented by FreeMorph [7] generate transitions by interpolating endpoint latents (e.g., via Slerp [7]) and injecting endpoint information through attention. While efficient, FreeMorph implicitly assumes that the endpoints are approximately *spatially aligned*.

However, this assumption does not always hold. When there exist large layout discrepancies between the input source and target images (as shown in Fig. 1), implicit methods cannot fully guarantee accurate spatial alignment between the two endpoint images. Accordingly, methods such as FreeMorph may produce distorted latent trajectories during latent interpolation, which further leads to ghosting artifacts, translucent overlays, and temporal incoherence in the intermediate images generated after diffusion denoising. Please refer to the image indicated by the arrow in Fig. 1 as an example.

To resolve this structural entanglement, the most intuitive remedy is to explicitly align the two endpoints before interpolation and mixing. However, achieving such effective alignment is non-trivial. Directly warping RGB images or relying on traditional low-level optical flow often fails, since the source and target typically share no pixel-level correspondence under severe layout discrepancies. To address this issue, we argue that the explicit alignment should satisfy two critical properties. **1): Semantic-aware alignment.** The alignment should be conducted at a higher semantic level rather than low-level pixels, so as to ensure structural coherence between the two input images. **2): Diffusion-compatible alignment.** Direct spatial warping on diffusion latents would break the inherent noise statistics required for stable denoising, resulting in unstable sampling and

distorted textures. Thus, the alignment mechanism must be naturally compatible with the diffusion model.

To satisfy these dual requirements, we decouple the spatial alignment from the generative refinement. Instead of forcing the diffusion model to implicitly resolve severe layout discrepancies during sampling, we establish a reliable geometric correspondence guided by semantic features beforehand. This naturally leads to our core design principle: transport geometry first, then denoise.Guided by this principle, we propose AlignMorph, a novel tuning-free diffusion-based image morphing framework that operates in two decoupled stages.

First, **Global Semantic Transport** establishes geometric alignment between endpoints. We estimate dense semantic correspondences via entropic optimal transport, derive reliability-gated warp fields, and apply a reliability-aware multiband latent transport that moves structural components while preserving stochastic variations. Together, these steps align endpoint representations into a shared coordinate frame in a diffusion-compatible manner, addressing large-displacement misalignment without corrupting noise statistics. The two images marked by green bounding boxes in Fig. 1 visualize the geometric alignment from source to target and vice versa, showing that our method achieves semantically consistent geometry alignment rather than naive pixel-level correspondence. Second, **Coordinate-Aligned Generation** performs diffusion sampling under explicit coordinate consistency. We employ a symmetric bi-phase attention handoff that switches between source-aligned and target-aligned memory banks at the morph midpoint, ensuring that queries and memories remain in the same coordinate frame throughout denoising and preventing cross-frame retrieval conflicts.

Extensive experiments on large-displacement morphing benchmarks show that AlignMorph consistently improves structural integrity and temporal smoothness over existing works, while requiring no further tuning. Please see Fig. 1 as a comparison example. In summary, our contributions are threefold:

- We propose AlignMorph, a tuning-free diffusion morphing framework based on a *transport-then-denoise* paradigm, which explicitly decouples geometric alignment from generative refinement.
- We propose Global Semantic Transport mechanism. It achieves reliable geometric alignment by leveraging optimal-transport-based dense semantic correspondence to drive a reliability-aware multiband latent transport, aligning unaligned latents into a shared coordinate frame without corrupting diffusion noise statistics.
- We propose a coordinate-aligned generation mechanism with a symmetric bi-phase handoff, ensuring that queries and memory features remain in the same coordinate frame throughout denoising.

2 Related Work

2.1 Image Morphing

Image morphing aims to generate smooth visual transitions between two endpoint images. Traditional methods [3, 4, 27, 53] typically decompose morphing into correspondence estimation and warping-based interpolation, but require manual intervention and struggle with complex appearance changes. Learning-based approaches [35,44,49] learn morphing transformations end-to-end but lack generalization to arbitrary image pairs.

Recent diffusion models [13, 17, 38, 42] enable high-quality generative morphing. Optimization-based methods [54, 55] achieve strong fidelity via per-pair tuning (e.g., LoRA [21]), but incur prohibitive computational costs. Tuning-free methods [7,26,52] eliminate this bottleneck through latent interpolation and attention feature injection, which implicitly assume spatial alignment between endpoints. Under severe spatial misalignment, this assumption fails: unaligned operations blindly entangle foreground and background semantics, causing ghosting and temporal incoherence [29]. We overcome this limitation by establishing explicit semantic correspondence via optimal transport prior to diffusion sampling, effectively decoupling geometric transport from generative refinement.

2.2 Geometric Alignment in Latent Space

Establishing correspondences across images has been extensively studied through optical flow [20, 22, 31, 47, 48], feature matching [2, 30], and semantic correspondence methods [9, 19, 32, 33, 50]. Recent works leverage self-supervised representations (e.g., DINO [8], DINOv2 [36]) enhanced with high-resolution feature upsampling [14] for robust matching despite appearance variations. Optimal transport (OT) [37] provides a principled framework for soft matching, with entropic regularization [12] enabling efficient computation. Several works [11, 28, 40, 41] have applied OT to vision tasks.

However, applying geometric transport to diffusion-based morphing introduces unique challenge. Diffusion latents encode both semantic structure (low-frequency) and stochastic noise (high-frequency) [1, 10, 46]. Naively warping the entire latent corrupts high-frequency noise distributions, leading to blurred textures and unstable sampling. While frequency-domain analysis has been explored in generative models [23–25], no prior work explicitly addresses the tension between geometric alignment and noise preservation in diffusion latent space. Our frequency-aware multiband latent transport resolves this by incorporating reliability-gated warping and spectral decomposition: we aggressively transport low-frequency structural components while selectively preserving high-frequency statistics, enabling geometric alignment without corrupting diffusion priors.

2.3 Attention Mechanisms for Generative Consistency

Attention mechanisms have become central to diffusion-based synthesis. Cross-attention enables conditional generation [39, 42, 43] by injecting textual or visual

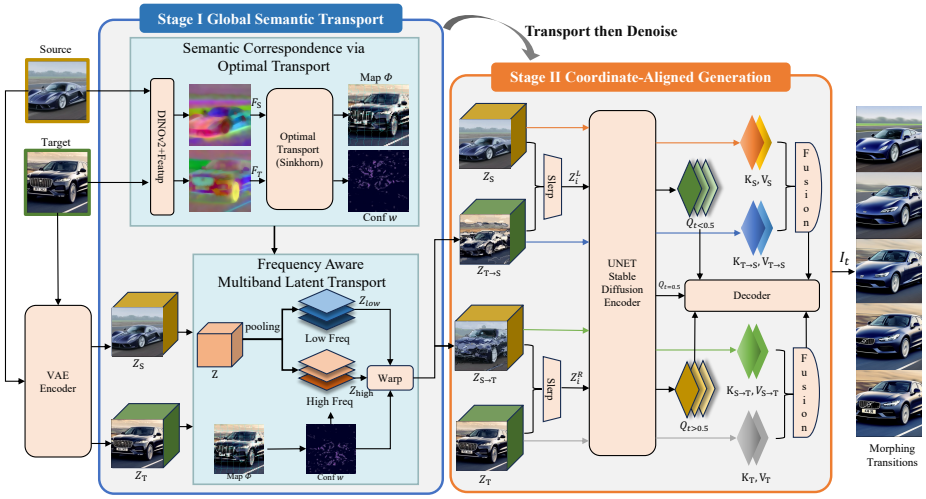

Fig. 2: Overview of the AlignMorph framework. Following a *transport-then-denoise* paradigm, our pipeline explicitly decouples geometric alignment from generative refinement. **Stage I** (left) establishes a shared coordinate frame: it first extracts Semantic Correspondence via Optimal Transport, and then applies Frequency-Aware Multiband Latent Transport to align endpoint latents without corrupting noise statistics. **Stage II** (right) executes Coordinate-Aligned Generation, utilizing a symmetric bi-phase attention handoff during UNet denoising to enforce geometric consistency and synthesize artifact-free transitions.

conditions into denoising. Recent works [5,34,56] extend this with spatial control signals (e.g., ControlNet [56]). For image editing, methods such as Prompt-to-Prompt [16], Plug-and-Play [51], and MasaCtrl [6] manipulate attention maps for tuning-free editing. For morphing, FreeMorph [7] adapts cross-attention using endpoint features as memory banks. However, under severe spatial misalignment, this induces structural entanglement: mixing unaligned memories blindly aggregates geometrically incompatible features, producing ghosting and incoherence. Video diffusion models [18, 45] incorporate temporal attention but assume short-range continuity, failing under such severe misalignment. To resolve this structural entanglement, our coordinate-aligned generation mechanism employs a bi-phase handoff that explicitly switches the coordinate frame at the temporal midpoint, ensuring queries and memories share a consistent geometric space throughout denoising.

3 Methodology

To synthesize a smooth intermediate sequence $\{I_t\}_{t\in(0,1)}$ between two unaligned endpoints S and T without per-pair optimization, AlignMorph introduces a novel *transport-then-denoise* paradigm that performs explicit semantic transport first

by establishing semantic-level correspondence between S and T, thereby guiding robust image morphing via diffusion denosing. Figure 2 illustrates the overall pipeline of AlignMorph, which comprises the following two stages. (i) **Global Semantic Transport**: We first obtain semantic-aware dense correspondences via optimal transport (Sec. 3.1). These correspondences are subsequently utilized to explicitly align the endpoint representations into a shared coordinate frame through a frequency-aware multiband latent transport (Sec. 3.2). (ii) **Coordinate-Aligned Generation** (Sec. 3.3): To generate structurally coherent intermediate transitions, this stage enforces geometric consistency throughout the diffusion denoising process via a symmetric bi-phase attention handoff.

3.1 Semantic Correspondence via Optimal Transport

Given two endpoint images S and T, the primary and crucial step is to establish the correspondence between the two images, which lays the foundation for subsequent alignment and morphing. As discussed earlier, since S and T often exhibit large layout discrepancies, this correspondence should be both *semantic-aware* and *diffusion-compatible*. Below, we shall gradually describe how we establish such correspondences to satisfy the above two requirements.

(1) For semantic-aware property. Our goal is to build correspondence at a high semantic level rather than the pixel level. To this end, we formulate the semantic correspondence as an Optimal Transport (OT) problem in the feature space. Specifically, as shown in the top-left of Fig. 2, we first extract dense semantic features $F_S \in \mathbb{R}^{H_S \times W_S \times d}$ and $F_T \in \mathbb{R}^{H_T \times W_T \times d}$ using a ViT encoder enhanced by FeatUp [15]. For computational efficiency, we compute the transport plan on a downsampled grid. Let $\{x_i\}_{i=1}^{N_S}$ and $\{y_j\}_{j=1}^{N_T}$ denote the resulting token sets, where $x_i, y_j \in \mathbb{R}^d$, $N_S = H_S W_S$, and $N_T = H_T W_T$. Each token is explicitly associated with its 2D spatial coordinate: $p_i \in \mathbb{R}^2$ for x_i and $q_j \in \mathbb{R}^2$ for y_j.

To establish a structurally coherent mapping between these unaligned feature sets, we formulate the dense matching task as the OT problem. Intuitively, we view the spatial tokens of each image as two sets of points with equal weight. Matching these images is equivalent to transporting the semantic features from the source to the target. The "transportation cost" for assigning a source token x_i to a target token y_j is determined by their semantic dissimilarity. We explicitly measure this using the cosine distance:

$$C_{ij} = 1 - \frac{\langle x_i, y_j \rangle}{\|x_i\|_2 \|y_j\|_2} \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product (dot product) of the two vectors, and $\| \cdot \|_2$ represents their L_2 norm (magnitude).

Consequently, finding a reliable dense correspondence simply translates to finding a global transport plan $T^\star \in \mathbb{R}_+^{N_S \times N_T}$ that minimizes the total transportation cost, ensuring that overall structural coherence is maintained without collapsing into many-to-one mapping errors. We solve this balanced entropic OT

problem using log-domain Sinkhorn iterations to obtain $T^\star$. Finally, to convert this dense probabilistic plan into a deterministic geometric warp field, we apply barycentric projection. Specifically, by row-normalizing the transport plan as $\bar{T}_{ij} = T^\star_{ij}/(\sum_k T^\star_{ik} + \delta)$, we define the forward coordinate map as:

$$\Phi_{S\to T}(p_i) = \sum_j \bar{T}_{ij}\, q_j \tag{2}$$

Here, $\Phi_{S\to T}$ is a coordinate map defined on the feature grid rather than a pixel-valued image, and the backward map $\Phi_{T\to S}$ is obtained analogously.

Note that, soft OT may be unreliable in ambiguous or occluded regions. To filter out unreliable matches, we compute a forward-backward consistency error:

$$\Delta(p_i) = \big\| p_i - \Phi_{T\to S}(\Phi_{S\to T}(p_i)) \big\|_2. \tag{3}$$

We then define a reliability weight $w(p_i) \in [0,1]$ by combining OT confidence and cycle consistency. Specifically, we use a plan sharpness score (row-max mass) to form a confidence value $c(p_i) \in [0,1]$, and apply a Gaussian penalty to the consistency error $\Delta(p_i)$:

$$w(p_i) = c_{S\to T}(p_i) \cdot c_{T\to S}(\Phi_{S\to T}(p_i)) \cdot \exp\Big(- \frac{\Delta(p_i)^2}{2\sigma^2} \Big). \tag{4}$$

(2) For diffusion-compatible property. Having established the semantic correspondence $\Phi_{S\to T}$ and its reliability weight w, we must appropriately project this mapping into the latent space. Specifically, considering that the diffusion sampling operates on a coarse latent grid of size $H_z \times W_z$, we thus project the model-space correspondence $\Phi_{S\to T}$ to this latent space by converting the coordinate map into a displacement (flow) field and resizing it. Let $r(p) = \Phi_{S\to T}(p) - p$ be the model-space displacement on the source grid. We bilinearly resize r to the latent resolution and rescale it to latent pixel units to obtain $r^z(u)$, yielding the latent warp operator and reliability mask:

$$\Phi^z_{S\to T}(u) = u + r^z(u), \qquad w^z(u) = \mathrm{Resize}(w)(u). \tag{5}$$

We compute the opposite-direction latent map and reliability mask, denoted as $\Phi^z_{T\to S}$ and $w^z_{T\to S}$, analogously. The forward mask is denoted as $w^z_{S\to T}$.

To validate the effectiveness of the proposed semantic alignment, we decode and visualize the directly warped latents, as illustrated in Fig. 3. As can be observed, our approach behaves exactly as theoretically expected: it successfully extracts robust, part-level semantic correspondences rather than superficial pixel matching. Although these directly decoded latents are not perfectly realistic, they reveal a crucial capability. For instance, in the vehicle examples, the transport plan explicitly maps specific structural components such as wheels to wheels and rear wings to rear wings, despite severe pose variations and distinct environmental contexts (e.g., the green versus blue race tracks). Similarly, fine-grained features (e.g., the dog's facial parts) are routed to their correct

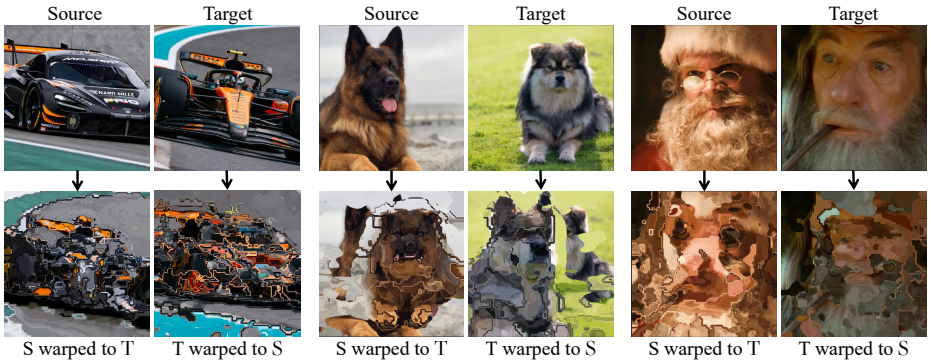

Fig. 3: Bidirectional semantic warping via optimal transport. Top row: Original source and target image pairs. Bottom row: Directly decoded warped latents.

anatomical positions. This robust, fine-grained structural alignment provides a geometrically coherent initialization, effectively resolving the spatial mismatch that fundamentally undermines traditional attention-based morphing.

3.2 Frequency-Aware Multiband Latent Transport

Having obtained the latent warp operator Φ^z and reliability mask w^z, the subsequent step is to utilize them to geometrically align the endpoint diffusion latents. The conventional approach is standard spatial warping, which treats the diffusion latent z as a monolithic signal and applies a single geometric transform to all components. However, a fundamental challenge arises: this rigid treatment fails to reconcile macroscopic structural alignment with the preservation of fine-grained details and noise statistics. While low-frequency components (capturing pose and layout) can be transported reliably, high-frequency components (capturing texture details and stochastic noise) are highly sensitive to misalignment. Forcing the same warp onto these high-frequency signals disrupts their local statistics, inevitably leading to texture degradation (e.g., blurring, broken patterns) and unstable denoising. To address this issue, we propose Frequency-Aware Multiband Latent Transport. Its core idea is to explicitly decouple geometric alignment from texture synthesis by aggressively transporting low-frequency structural components while selectively preserving high-frequency noise.

Specifically, we decompose the latent tensor z into a low-frequency structural component and a high-frequency residual:

$$z = z_{\text{low}} + z_{\text{high}}. \tag{6}$$

We obtain z_{low} via spatial average pooling (kernel size k=9, stride 1 with replicate padding), which suppresses high-frequency variations while preserving coarse semantic layout. The residual

$$z_{\text{high}} = z - z_{\text{low}} \tag{7}$$

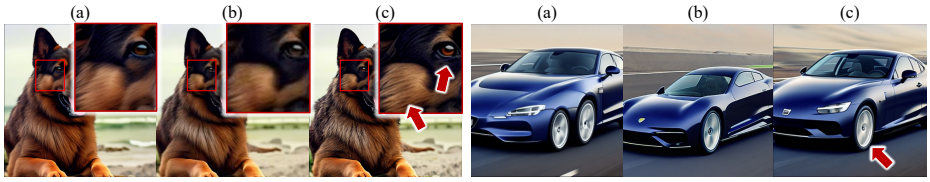

Fig. 4: Comparison among (a) FreeMorph [7], (b) full latent warping (Warp-All), and (c) our multiband transport on two examples.

encodes fine-grained textures and stochastic noise variations that should not be aggressively warped in unreliable regions.

Let Φ^z denote the projected latent-grid warp field (Sec. 3.1), and let $w^z \in [0,1]$ be the corresponding reliability weight on the latent grid obtained from Eq. (5). We denote by $\mathcal{W}(\cdot;\Phi^z)$ a warping operator that maps each latent location according to Φ^z. Our transported latent z' is defined as

$$z' = \underbrace{\mathcal{W}(z_{\text{low}};\Phi^z)}_{\text{Geometric guidance}} + \underbrace{w^z \cdot \mathcal{W}(z_{\text{high}};\Phi^z) + (1-w^z)\cdot z_{\text{high}}}_{\text{Statistical preservation}}. \tag{8}$$

Using Eq. (8), we construct geometrically aligned endpoint latents:

$$z_{T\to S} = \mathcal{T}(z_T;\Phi^z_{T\to S}, w^z_S), \qquad z_{S\to T} = \mathcal{T}(z_S;\Phi^z_{S\to T}, w^z_T), \tag{9}$$

where $\mathcal{T}(\cdot)$ denotes the multiband transport operator in Eq. (8), and w^z_S, w^z_T are the reliability weights derived from cycle-consistency on the corresponding coordinate frames.

To validate the necessity of our frequency-aware multiband transport, we analyze two representative morphing scenarios in Fig. 4: one where FreeMorph produces reasonable transitions (Left), and another where it fails completely with severe ghosting artifacts (Right). In both cases, applying full latent warping (Warp-All) indiscriminately alters the latent representation. This rigid operation disrupts the local high-frequency statistics essential for detail synthesis, causing blurred fine-grained textures (e.g., the dog's fur) and destroying specific semantic details (e.g., car headlights and sharp body edges). By explicitly decoupling frequencies, our multiband transport selectively aligns the low-frequency structural layout while strictly preserving the original high-frequency noise. Consequently, our approach eliminates ghosting where baselines fail, and maintains sharp visual fidelity that even exceeds the baseline in successful cases.

3.3 Coordinate-Aligned Generation

With the obtained geometrically aligned endpoint latents $z_{T\to S}$ and $z_{S\to T}$ via multiband transport, the next step is to synthesize the intermediate transition frames without disrupting this structural consistency during diffusion denoising. The conventional approach in existing tuning-free methods is to interpolate unaligned endpoint latents to form the intermediate query $Q^l_{i,k}$, and directly mix

unaligned endpoint memories to form the K/V branches. However, this approach forces the attention mechanism to aggregate features across disparate geometric layouts, inevitably causing severe structural entanglement even before the attention is calculated. To address this issue, we propose Coordinate-Aligned Generation. Its core idea is to explicitly construct both the interpolated query and the mixed endpoint memories within a phase-consistent coordinate frame, thereby guaranteeing that spatial indices consistently refer to the same semantic structures and making the attention fusion geometrically coherent.

Specifically, let N be the number of frames and $i \in \{0, \dots, N-1\}$ the frame index. We denote morph progress by $\alpha_i = \frac{i}{N-1}$. Let $k \in \{1, \dots, K\}$ be the diffusion step, and $z_S, z_T \in \mathbb{R}^{C \times H_z \times W_z}$ be endpoint latents. Using the latent-grid transport results from Eq. (9), we build two directional latent paths via spherical linear interpolation (Slerp) from [7]:

$$z_i^L = \mathrm{Slerp}(z_S, z_{T\to S}, \alpha_i), \qquad z_i^R = \mathrm{Slerp}(z_{S\to T}, z_T, \alpha_i). \tag{10}$$

Let $m = \lfloor N/2 \rfloor$ be the midpoint index. To reduce discontinuity at the frame handoff, we map the right midpoint back to the source frame with a weaker transport and fuse the two midpoint candidates, $\mathcal{T}_{1/2}$ denotes half-strength transport:

$$\tilde{z}_m^{R\to S} = \mathcal{T}_{1/2}(z_m^R; \Phi^z_{T\to S}, w_S^z), \qquad z_m = \mathrm{Slerp}(z_m^L, \tilde{z}_m^{R\to S}, 0.5). \tag{11}$$

The initialized sequence is

$$z_i = \begin{cases} z_i^L, & i < m, \\ z_m, & i = m, \\ z_i^R, & i > m. \end{cases} \tag{12}$$

For an attention layer l at diffusion step k, the query $Q_{i,k}^l$ is computed from the current frame latent z_i via the UNet hidden state at that layer. Hence, $Q_{i,k}^l$ is expressed in the same spatial coordinate frame as the current morphing latent. The key and value memories are constructed from two endpoint latents of the current phase. Given an endpoint pair $(z^{(1)}, z^{(2)})$, we form $(K_{i,k}^{(1)}, V_{i,k}^{(1)})$ and $(K_{i,k}^{(2)}, V_{i,k}^{(2)})$ from their corresponding UNet features at layer l and step k. The attention output is then interpolated as:

$$\mathrm{Out}_{i,k}^l = (1-\lambda_i)\,\mathrm{Attn}(Q_{i,k}^l, K_{i,k}^{(1)}, V_{i,k}^{(1)}) + \lambda_i\,\mathrm{Attn}(Q_{i,k}^l, K_{i,k}^{(2)}, V_{i,k}^{(2)}), \tag{13}$$

where $\lambda_i \in [0,1]$ is a monotone frame-wise schedule with $\lambda_0 = 0$ and $\lambda_{N-1} = 1$.

To explicitly keep the query and memory in the same coordinate frame, inference is performed in two phases by instantiating $(z^{(1)}, z^{(2)})$ as follows: For early frames $(i \le m)$, $(z^{(1)}, z^{(2)}) = (z_S, z_{T\to S})$, aligning the memory with the source frame; for late frames $(i > m)$, $(z^{(1)}, z^{(2)}) = (z_{S\to T}, z_T)$, aligning the memory with the target frame. This phase-wise construction ensures that Q and both K/V branches are always defined in a consistent geometry.

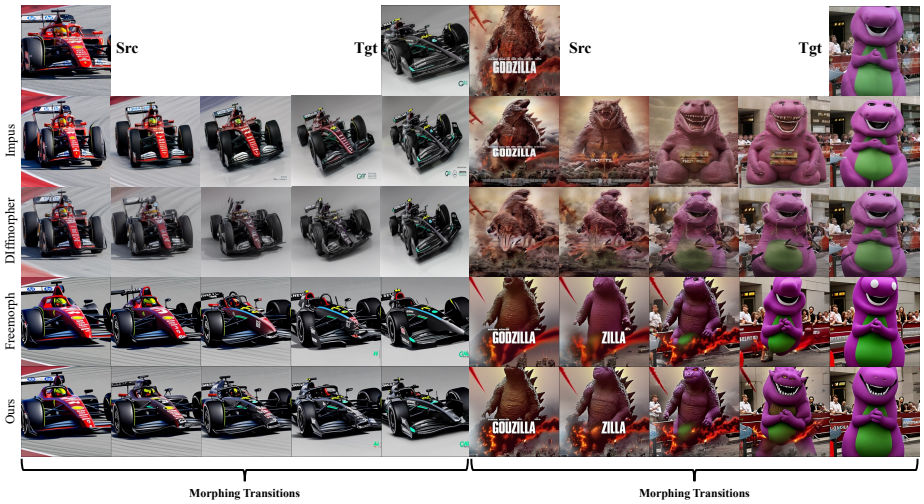

Fig. 5: Qualitative comparison. Left (F1 car), Right (Godzilla).

4 Experiments

4.1 Setup and Implementation Details

Implementation Details. We implement AlignMorph on Stable Diffusion v2.1 using a single RTX 3090. Strictly following the FreeMorph protocol for fair comparison, we use a DDIM scheduler ($K = 50$, edit strength 0.8) for 40-step inversion and denoising. We generate $N = 7$ frames at 768×768 resolution with a CFG scale of 7.5, keeping all unmentioned configurations (e.g., seeds, noise initialization) identical to the baseline. For global semantic transport, DINOv2+FeatUp features are extracted and adaptively downsampled to $\leq 16,384$ tokens for efficiency. We solve entropic optimal transport via log-domain Sinkhorn iterations ($\varepsilon = 0.1$, 100 steps). Reliability masks use a cycle-consistency threshold $\tau = 8$ pixels and a minimum confidence of 0.2. In multiband transport, low-frequency components are extracted via average pooling ($k = 9$, stride 1, padding 4). The latent warping operates via bilinear grid sampling with border padding.

4.2 Quantitative Evaluations

Following prior work [7, 54, 55], we evaluate fidelity and transition quality using three metrics: (1) **FIDmean**, (2) **LPIPSsum** between adjacent frames, and (3) **PPLsum** between adjacent frames. Lower is better for all metrics.

Table 1 reports the results on MorphBench and Morph4Data. AlignMorph achieves superior performance compared with both optimization-based and tuning-free baselines. In addition to quality metrics, we compare runtime efficiency. AlignMorph requires 115s per image pair on a single RTX 3090 GPU, including 25s for offline correspondence and 90s for diffusion sampling. This runtime is

Table 1: Quantitative comparison on MorphBench and Morph4Data. Results for the first three baselines are directly cited from [7]. For rigorous benchmarking and to ensure a strictly fair comparison, we additionally report the reproduced results of FreeMorph using their official implementation. ↓ indicates lower is better.

	MorphBench			Morph4Data			Overall		
Method	LPIPS↓	FID↓	PPL↓	LPIPS↓	FID↓	PPL↓	LPIPS↓	FID↓	PPL↓
IMPUS [54]	130.52	152.43	3263.03	134.88	210.66	3199.90	265.40	174.76	6462.93
DiffMorpher [55]	90.57	157.18	2264.20	98.56	292.54	2394.05	189.13	209.10	4658.25
FreeMorph [7]	84.91	141.32	2122.80	80.30	201.09	2007.52	165.21	152.88	4130.32
FreeMorph (Reproduced)	85.01	139.79	2145.37	79.56	203.32	1993.42	164.57	154.45	4138.79
AlignMorph (ours)	**79.55**	**131.98**	**2010.63**	**74.23**	**188.44**	**1896.30**	**153.78**	**140.12**	**3896.93**

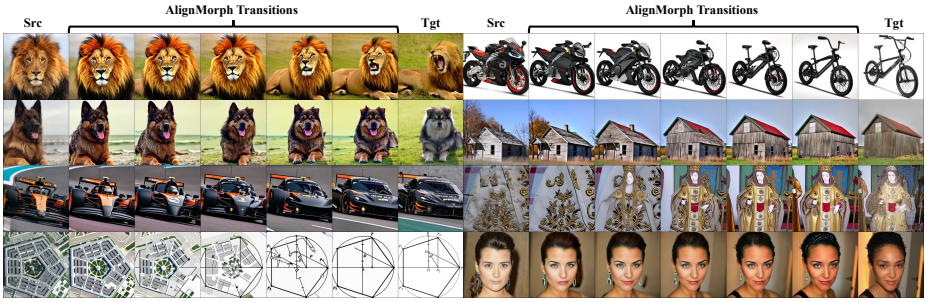

Fig. 6: More transitions generated by AlignMorph. Our method produces robust transitions across diverse image pairs under different scenarios.

comparable to other tuning-free methods such as FreeMorph (90s) under the same sampling schedule. In contrast, optimization-based approaches (e.g., IMPUS and DiffMorpher) require per-pair fine-tuning or latent optimization, typically taking tens of minutes per pair. AlignMorph avoids any optimization and achieves approximately 20×–50× speedup over optimization-based methods.

User Study. To evaluate perceptual quality, we conducted a user study with 25 participants, ranging from university students to AI researchers. Each participant was presented with 50 randomly selected test cases and asked to choose the best result based on structural integrity and smoothness. As shown in Table 3, AlignMorph outperforms all baselines with a preference rate of **50.9%**, confirming its superiority in generating natural and consistent transitions.

4.3 Qualitative Evaluations

Qualitative Comparisons. As shown in Fig. 5, we evaluate AlignMorph against baselines under large layout discrepancies. Optimization-based methods (IMPUS, DiffMorpher) struggle with structural consistency, producing non-smooth transitions and hallucinating unrealistic artifacts entirely absent from the inputs. While the tuning-free baseline FreeMorph mitigates these severe hallucinations, it still suffers from its implicit spatial alignment. Specifically, it exhibits temporally inconsistent color shifts and structural distortions in the F1 car (left). For

Table 2: Ablation study. Lower is better.

Variant	LPIPS↓	FID↓	PPL↓
FreeMorph	80.30	201.09	2007.52
Random OT	132.14	297.60	2803.76
w/o Multiband	88.16	231.97	2255.20
w/o Handoff	91.70	257.91	2670.11
w/o KV warp	76.12	193.76	1940.46
AlignMorph	**74.23**	**188.44**	**1896.30**

Table 3: User Preference Rate (%).

Method	DiffMorpher	FreeMorph	IMPUS	**Ours**
Preference	8.2	26.4	14.5	**50.9**

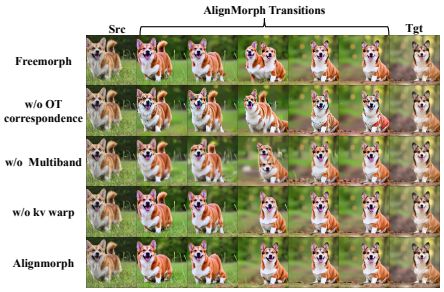

Fig. 7: Qualitative comparison. AlignMorph preserves structural integrity better than baselines.

the Godzilla morph (right), FreeMorph loses high-frequency details, resulting in blurry textures and intermediate frames with collapsed, meaningless semantics. In contrast, by explicitly aligning geometry prior to denoising, AlignMorph successfully maintains crisp textures, coherent structural boundaries, and semantically meaningful intermediate states without any blending artifacts.

Qualitative Results. We show additional morphing sequences generated by AlignMorph in Fig. 6. Across diverse scenes and endpoint misalignment, AlignMorph produces coherent object transport and temporally stable synthesis.

4.4 Ablation Study

This section analyzes the effects of AlignMorph components and connects them to the failure modes of tuning-free interpolation under large deformation. We conduct ablation experiments on Morph4Data to analyze the contribution of each component in AlignMorph. Results are summarized in Table 2.

w/o semantic correspondence. Replacing OT-based matching with a random mapping (20% random displacement) degrades all metrics (Table 2). Visually (Fig. 7), it introduces severe scattered noise and dark artifacts. This occurs because semantically unaware mapping blindly warps incompatible background features into the foreground, corrupting local latent statistics and proving the necessity of accurate semantic transport.

w/o multiband latent transport. Removing the frequency-aware decoupling and directly warping the entire latent space results in a noticeable performance drop. Visually, also(Fig. 7), textures become visibly blurred, and fine details in challenging regions (e.g., fur) are severely weakened. This validates the necessity of preserving high-frequency noise statistics. (We also verified that rigorous frequency decoupling via Fast Fourier Transform yields comparable generative quality, justifying our choice of spatial average pooling for its computational efficiency and simplicity.)

w/o bi-phase coordinate handoff. In this variant, we remove the two-phase coordinate schedule and keep the entire morphing process in a single co-

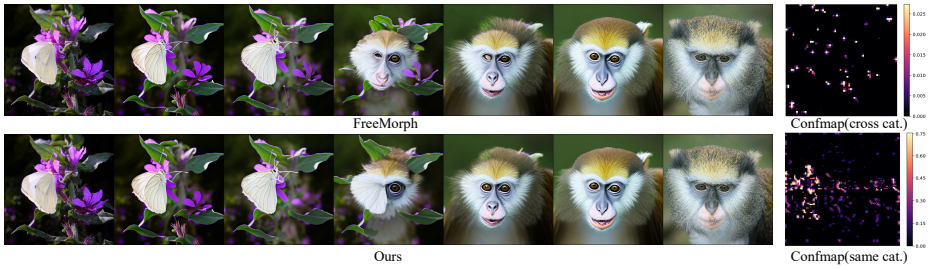

Fig. 8: Failure analysis on extreme cross-category morphing. Left: comparison with FreeMorph on a flower-to-monkey pair. Right: confidence maps for cross-category and same-category pairs.

ordinate frame. Specifically, the target representation is continuously warped into the source coordinate system, and no world-frame switch is performed at the morph midpoint. As a result, the trajectory remains constrained within the source-aligned geometry and cannot fully transition into the target coordinate frame. The proposed bi-phase handoff resolves this issue by explicitly switching the world coordinate frame at the midpoint, allowing the trajectory to progressively adopt the target geometry while maintaining consistent transitions.

Empirically, we observe that warping the *UNet input latent* (from which the query Q is derived) serves as the primary driver for geometric alignment. Even when endpoint memories (K, V) remain unwarped (**w/o KV warp**), our bi-phase handoff framework still generates structurally coherent transitions. This behavior aligns with the fundamental nature of attention as a coordinate-sensitive retrieval operator: Q dictates *where* to retrieve, while (K, V) dictate *what* is retrieved. Under large deformations, the dominant failure mode is a *query-side spatial mismatch*—queries are spatialized on an unaligned intermediate grid, whereas endpoint memories reside in a fixed endpoint frame. Warping the input latent explicitly aligns the *query-side coordinate frame* prior to $QK^\top$ matching, fundamentally resolving severe structural entanglement. Subsequently, aligning (K, V) ensures the *content consistency* of the retrieved features against the transported geometry, effectively eliminating residual texture drift and minor ghosting. Thus, while Q alignment establishes the macroscopic geometric structure, KV alignment provides a vital complementary refinement that maximizes fine-grained texture fidelity and overall visual quality.

5 Limitations and Future Work

Current morphing benchmarks primarily focus on image pairs with partial semantic or structural overlap, where meaningful correspondence and geometric alignment are well-defined. As shown in Fig. 8, extreme cross-category pairs instead expose a different regime: the semantic correspondence itself becomes invalid. For example, in flower-to-monkey morphing, there is no canonical map-

ping between petals, leaves, and facial parts. This is reflected by the confidence map, where the cross-category case produces consistently low responses. In contrast, when the two endpoints belong to the same category but undergo large spatial displacement, the confidence map remains much stronger, indicating that valid part-level correspondence can still be established. Accordingly, our reliability gate suppresses unreliable transport in the former case rather than enforcing arbitrary OT warps. AlignMorph therefore degrades toward an interpolation-dominant trajectory, which avoids additional wrong-warp artifacts but also reduces the advantage of explicit alignment over interpolation-based transitions.

This limitation is different from large deformation with valid semantic correspondence. For same-category pairs, the confidence map remains much denser, indicating that reliable part-level correspondence can still be established despite large layout displacement. Therefore, the main boundary of AlignMorph is not deformation magnitude, but the existence of meaningful semantic correspondence. Extending morphing beyond shared structural similarity requires redefining the alignment objective and perceptual criteria for open-category generative blending, which we leave for future work.

6 Conclusion

In this paper, we presented **AlignMorph**, a novel tuning-free diffusion morphing framework that tackles the severe structural entanglement prevalent under large layout discrepancies. Guided by the *transport-then-denoise* paradigm, we explicitly decouple geometric alignment from generative refinement. Specifically, our framework first employs Global Semantic Transport to establish dense semantic correspondences via optimal transport, which are then projected into the latent space through a frequency-aware multiband transport. This crucial step guarantees reliable macro-geometric alignment while strictly preserving the high-frequency noise statistics inherent to diffusion priors. Subsequently, Coordinate-Aligned Generation synthesizes the intermediate transitions via a symmetric bi-phase attention handoff, ensuring strict spatial consistency throughout the denoising process. Extensive experiments demonstrate that AlignMorph effectively eliminates ghosting artifacts and achieves superior performance without the need for expensive per-pair optimization. We hope this work highlights the necessity of explicit geometric alignment in diffusion-based generation and inspires future research into more robust generative transport mechanisms.

Acknowledgements

This work was supported by the National Key Research and Development Program for Young Scientists of China (No. 2024YFB3310100), the Natural Science Foundation of Hubei Province (No. 2025AFB592), and the Natural Science Foundation of Wuhan (No. 2025040601020216).

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
3. Beier, T., Neely, S.: Feature-based image metamorphosis. Proceedings of the 19th annual conference on Computer graphics and interactive techniques (1992), `https://api.semanticscholar.org/CorpusID:9124441`
4. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence **11**(6), 567–585 (2002)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
6. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22560–22570 (2023)
7. Cao, Y., Si, C., Wang, J., Liu, Z.: Freemorph: Tuning-free generalized image morphing with diffusion model. arXiv preprint arXiv:2507.01953 (2025)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
9. Cho, S., Hong, S., Jeon, S., Lee, Y., Sohn, K., Kim, S.: Cats: Cost aggregation transformers for visual correspondence. Advances in Neural Information Processing Systems **34**, 9011–9023 (2021)
10. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
11. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(9), 1853–1865 (2017)
12. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems. vol. 26 (2013)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
14. Fu, S., Hamilton, M., Brandt, L., Feldman, A., Zhang, Z., Freeman, W.T.: Featup: A model-agnostic framework for features at any resolution (2024)
15. Fu, S., Hamilton, M., Brandt, L., Feldman, A., Zhang, Z., Freeman, W.T.: Featup: A model-agnostic framework for features at any resolution (2024), `https://arxiv.org/abs/2403.10516`
16. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
18. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 8633–8646 (2022)

19. Hong, S., Cho, S., Nam, J., Lin, S., Kim, S.: Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In: European Conference on Computer Vision. pp. 108–126 (2022)
20. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial Intelligence **17**(1-3), 185–203 (1981)
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
22. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9772–9781 (2021)
23. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Advances in Neural Information Processing Systems. vol. 34, pp. 852–863 (2021)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
25. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
26. Kye, D., Sung, J., Jeon, M., Oh, J.: Chimera: Adaptive cache injection and semantic anchor prompting for zero-shot image morphing with morphing-oriented metrics. arXiv preprint arXiv:2512.07155 (2025)
27. Lee, S., Wolberg, G., Shin, S.: Scattered data interpolation with multilevel b-splines. IEEE Transactions on Visualization and Computer Graphics **3**(3), 228–244 (1997)
28. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 978–994 (2010)
29. Liu, X., Li, Y., He, Q., Zhu, J., Ji, W., Yao, A., Zhu, J.: Interp3d: Correspondence-aware interpolation for generative textured 3d morphing. arXiv preprint arXiv:2601.14103 (2026)
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
31. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI'81: 7th international joint conference on Artificial intelligence. vol. 2, pp. 674–679 (1981)
32. Min, J., Lee, J., Ponce, J., Cho, M.: Hyperpixel flow: Semantic correspondence with multi-layer neural features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3395–3404 (2019)
33. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019)
34. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
35. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)
36. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

37. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
38. 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. arXiv preprint arXiv:2307.01952 (2023)
39. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
40. Rocco, I., Arandjelović, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6148–6157 (2017)
41. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
43. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems. vol. 35, pp. 36479–36494 (2022)
44. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation **32** (2019)
45. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
46. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
47. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8934–8943 (2018)
48. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419 (2020)
49. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. arXiv preprint arXiv:2104.15069 (2021)
50. Truong, P., Danelljan, M., Timofte, R.: Glu-net: Global-local universal network for dense flow and correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6258–6268 (2020)
51. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
52. Wang, C.J., Golland, P.: Interpolating between images with diffusion models. arXiv preprint arXiv:2307.12560 (2023)
53. Wolberg, G.: Image morphing: a survey. The visual computer **14**(8), 360–372 (1998)
54. Yang, Z., Yu, Z., Xu, Z., Singh, J., Zhang, J., Campbell, D., Tu, P., Hartley, R.: Impus: Image morphing with perceptually-uniform sampling using diffusion models. In: International Conference on Learning Representations (2023)

55. Zhang, K., Zhou, Y., Xu, X., Dai, B., Pan, X.: Diffmorpher: Unleashing the capability of diffusion models for image morphing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
56. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)