Versatile Transition Generation with Image-to-Video Diffusion

Zuhao Yang[1][*]   Jiahui Zhang[1]   Yingchen Yu[2]   Shijian Lu[1][†]   Song Bai[2]

[1]Nanyang Technological University   [2]ByteDance Inc.

https://mwxely.github.io/projects/yang2025vtg/index

Figure 1. Illustration of *Versatile Transition Generation* (VTG). VTG is capable of performing four types of transition generation, namely, (a) object morphing, (b) motion prediction, (c) concept blending, and (d) scene transition, within a single uniform framework.

(a) "a dog facing forward" → "a dog turning his head to the right"

(b) "a woman riding a horse" → "a woman moving forward on a horse"

(c) "an airplane" → "a cruise"

(d) "a wooden house in the forest" → "a wooden house in the snow"

Abstract

*Leveraging text, images, structure maps, or motion trajectories as conditional guidance, diffusion models have achieved great success in automated and high-quality video generation. However, generating smooth and rational transition videos given the first and last video frames as well as descriptive text prompts is far underexplored. We present **VTG**, a **V**ersatile **T**ransition video **G**eneration framework that can generate smooth, high-fidelity, and semantic-coherent video transitions. VTG introduces interpolation-based initialization that helps preserve object identity and handle abrupt content changes effectively. In addition, it incorporates dual-directional motion fine-tuning and representation alignment regularization that mitigate the limitations of the pre-trained image-to-video diffusion models in motion smoothness and generation fidelity, respectively. To evaluate VTG and facilitate future studies on unified transition generation, we collected **TransitBench**, a com-prehensive benchmark for transition generation that covers two representative transition tasks including concept blending and scene transition. Extensive experiments show that VTG achieves superior transition performance consistently across the four tasks.*

1. Introduction

The success of diffusion models [18, 43] in image synthesis [30, 34, 36, 38] has inspired a number of studies on diffusion-based video synthesis [19, 47, 58, 60]. Leveraging textual prompts, video frames, structure maps and even motion patterns, several studies [7, 41, 44, 48] have demonstrated impressive synthesis performance by generating realistic and high-fidelity videos automatically. Nevertheless, despite its significant value in various real-world tasks such as video and film production, generating high-quality transition videos conditioned on the first and last frames together with related text prompts remains largely underexplored.

Generating realistic transition videos is a non-trivial task. A high-quality transition generator should meet at least four

---

criteria: 1) semantic similarity to the input frames; 2) high fidelity to the input frames; 3) smoothness across the generated frames; and 4) alignment with the provided text prompts. Moreover, the research community in transition generation mostly relies on self-collected and well-filtered videos that are not accessible to the public [8, 54]. This further hinders the advancement of the research in transition video generation.

Most existing studies tackle the challenge of transition generation via two representative approaches. The first focuses on morphing, given two images of topologically similar objects. Most recent methods [52, 56] harness various deep interpolation techniques for generating rational object-level transition. However, they generate intermittent images rather than temporally coherent video frames, losing the transition smoothness especially while handling moving objects. The second focuses on video frame interpolation. Most existing studies attempt to estimate intermediate optical flows [25, 27, 35, 55], or leverage frame conditioning at training time [4, 5, 22, 50]. However, they tend to generate irrational object transitions with abrupt content changes, struggle with producing long transition sequences, and require time-consuming training on large-scale motion videos. On top of the above, existing studies either employ deep interpolation for conceptual blending transitions [16, 51] or incorporate a random-mask condition layer for scene transition [8], lacking a uniform framework that can work for multiple transition generation tasks. This prompts a compelling inquiry: *can we design a versatile transition generator that can handle various transition tasks with minimal adaptation across tasks?*

As illustrated in Figure 1, we propose four transition tasks to establish *versatile transition generation*, based on the types of input frames: (1) Object Morphing: The input frames are either the same object with different postures or different objects, as long as they are topologically similar; (2) Concept Blending: The input frames contain conceptually different objects (e.g., 'an airplane' and 'a cruise'); (3) Motion Prediction: The input frames can be seen as two moments in a video containing one (or more) moving object; (4) Scene Transition: The input frames are conceptually related scene images yet they either belong to two different domains (e.g., 'a wooden house in the forest' and 'a wooden house in the snow') or represent two distinct components of a scene (e.g., 'erupting volcano' and 'hot lava').

We design VTG, a versatile transition generation framework built upon image-to-video diffusion models. VTG features three designs for generating smooth, high-fidelity, and semantic-coherent transitions. First, we employ interpolation-based initialization that effectively mitigates abrupt content changes while handling input frames with significantly different contents. Specifically, we spherically interpolate latent Gaussian noises of the two input frames

and text embeddings of corresponding transition captions and leverage two LoRA-integrated U-Nets [20] to capture object-level semantics in the denoising steps. Second, we adopt a lightweight fine-tuning strategy that merges the predicted forward and backward noises which greatly improves the transition smoothness in motion predictions. The bidirectional sampling can be achieved by fine-tuning the pre-trained U-Net with a small collection of videos. Third, we introduce regularization with self-supervised visual encoder for video diffusion models to explicitly induce feature alignment, which helps learn meaningful representations and enhance the fidelity of generated transition videos. These components are logically cohesive and, in combination, distinguish VTG from prior work by enabling a single framework to handle diverse transition tasks with minimal task-specific adjustments. In addition, we collected *TransitBench*, a new benchmark of 200 pairs with first and last frames for concept blending and scene transition.

The contributions of this work can be summarized in three aspects. First, we propose a unified task of *versatile transition generation*, aiming at smooth and rational transition for object morphing, concept blending, motion prediction, and scene transition. Second, we design VTG, a novel and versatile framework that can generate semantic-relevant, high-fidelity and temporally coherent video transitions effectively. Third, we introduce TransitBench, a curated dataset for benchmarking concept blending and scene transition. With TransitBench and other public benchmarks, we demonstrate, both qualitatively and quantitatively, that VTG outperforms the state-of-the-art consistently across the four transition generation tasks.

2. Related Work

Image Morphing. Image morphing has been a long-standing problem in the computer graphics community [1, 49, 61]. Conventional morphing techniques either utilize correspondence-driven bidirectional image warping [2, 3, 9, 26] or minimize the path energy on a Riemannian manifold [14, 29, 33, 40] to obtain image transitions. While producing seamless morphing, they either require massive human involvement or fail to create new content beyond the two given images. Recently, several studies [52, 56] explored the prior knowledge in image diffusion models for real-world morphing with more object categories. Differently, we exploit a pretrained video diffusion model for image morphing, naturally enabling coherent transitions across frames.

Video Frame Interpolation. Video frame interpolation (VFI) has been extensively studied in the computer vision research community [11]. Classic VFI algorithms are closely related to optical flow prediction, which obtain interpolation by either forward warping [31] or backward

warping [21, 23, 25] based on the estimated flows. Several follow-up studies [22, 27, 35] focus on large motion problem in VFI. Recently, a line of research [12, 46] explores bounded video generation with image-to-video diffusion models under the guidance of the start and end frames. This approach fuses outputs from forward and backward paths by interpolating the predicted forward noise and its reversed counterpart. While generating coherent motion, it overlooks the identity preservation and struggles to produce natural transitions when the two input frames differ significantly in content.

Transition Generation. Transition generation is a far underexplored problem, largely due to its open-ended nature. Typically, a transition generation framework leverages two input frames (i.e., the first and last frames) and a transition caption to produce a scene-level transition video that connects two distinct narrative moments. This has been explored in [8] that incorporates random masking to selectively suppress information from the original latent code, allowing capturing transition subtleties between frames. In addition, [50, 54] inject the first and last frames as additional conditions by concatenating them with noisy latent codes along the channel dimension. Recently, a concurrent study [57] explores transition video generation with image-to-video diffusion as well. Differently, we formulate the task by *versatile transition generation* by unifying four transition tasks under the same framework, which greatly expands the applicability and significantly enhances the quality of the generated transition videos.

3. Method

3.1. Preliminaries

Latent diffusion models [36] are text-to-image diffusion models that operate in a compressed latent space. During forward diffusion process, it encodes an image sample $\mathbf{x}_0$ into a latent code $\mathbf{z}_0$ and corrupts $\mathbf{z}_0$ by Gaussian noise at each time step $t = 1, \ldots, T$:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

where β_t denotes the noise strength at time step t and $\mathbf{I}$ denotes the identity matrix with the same dimensions as $\mathbf{z}_0$. Given a text embedding $\mathbf{c}$, the backward denoising process can be parameterized as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, \mathbf{c}, t), \Sigma_\theta(\mathbf{z}_t, \mathbf{c}, t)), \quad (2)$$

where $\mu_\theta(\mathbf{z}_t, \mathbf{c}, t)$ is output by a U-Net [37], and $\Sigma_\theta(\mathbf{z}_t, \mathbf{c}, t)$ is determined by a noise scheduler (e.g., DDPM [18] or DDIM [42]). The model learns to predict noise via a parameterized noise estimator ϵ_θ under the guidance of an objective function:

$$\mathcal{L}_t = \mathbb{E}_{t\sim[1,T],\mathbf{z}_0,\mathbf{c},\epsilon_t\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \parallel \epsilon_t - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t) \parallel_2^2. \quad (3)$$

Our framework is built upon pretrained image-to-video diffusion models that incorporate temporal convolutions and temporal attention layers on top of latent diffusion models to establish temporal correlations between video frames. Given a video $\mathbf{x}_0 \in \mathbb{R}^{N \times 3 \times H \times W}$, where N is the number of frames and $H \times W$ indicates the image resolution, each frame is first encoded into a latent representation $\mathbf{z}_0 \in \mathbb{R}^{N \times C \times h \times w}$ via the VAE encoder $\mathcal{E}_v$. Both forward diffusion process and backward denoising process are then performed in the latent space. The generated video frames can be collectively obtained through the decoder $\mathcal{D}_v$. To inject image conditions, we fuse and incorporate text-conditioned and image-conditioned features by projecting the conditional image into a text-aligned embedding space. Meanwhile, the conditional image is concatenated with the initial per-frame noise to preserve more visual details.

3.2. Interpolation-based Initialization

In diffusion models, the initial Gaussian noise determines the coarse structure that emerges in the early denoising steps, while high-frequency details are refined later. Since existing image-to-video diffusion models [8, 50] randomly initialize the latent code $\mathbf{z}_0$ at inference time, each intermediate frame follows a distinct stochastic trajectory: colors, micro-textures, or even object pose shift slightly from frame to frame, yielding perceptual "flicker" [10].

To suppress this drift, we correlate the frame-wise latents by interpolating between the endpoint noises instead of resampling them independently. Prior work [6] indicates that linear interpolation often yields intermediate latent norms that are extremely unlikely under a unit Gaussian, whereas spherical linear interpolation (SLERP) preserves the Euclidean norm and keeps samples on-distribution. Therefore, we leverage SLERP between the latent noises of two input frames $\mathbf{z}_{t1}$ and $\mathbf{z}_{tN}$ as follows:

$$\mathbf{z}_{tn} = \frac{\sin((1 - \lambda_{noise})\phi)}{\sin\phi}\mathbf{z}_{t1} + \frac{\sin(\lambda_{noise}\phi)}{\sin\phi}\mathbf{z}_{tN}, \quad (4)$$

where $\phi = \arccos\left(\frac{\mathbf{z}_{t1}^T \cdot \mathbf{z}_{tN}}{\|\mathbf{z}_{t1}\|\|\mathbf{z}_{tN}\|}\right)$, and $\lambda_{\text{noise}} \in [0,1]$ denotes the parameter for latent interpolation. As shown in Figure 2, we concatenate $\mathbf{z}_{tn}$ with $\mathbf{z}_{t1}$ and $\mathbf{z}_{tN}$ along the frame dimension and inject them as the initial latent noise for DDIM sampling. It is worth noting that we only inject the interpolated latent noises at early denoising steps to preserve the appearance and motion priors from the image-to-video backbone as much as possible. With such meticulously designed latent injection strategy, our approach addresses the issue of random and abrupt transition frames.

Unlike GANs [15], whose latent space is structured and semantically meaningful, diffusion models operate in an unstructured noise space that does not explicitly encode high-level semantics [56]. This lack of structure can lead to artifacts when interpolating between semantically different in-
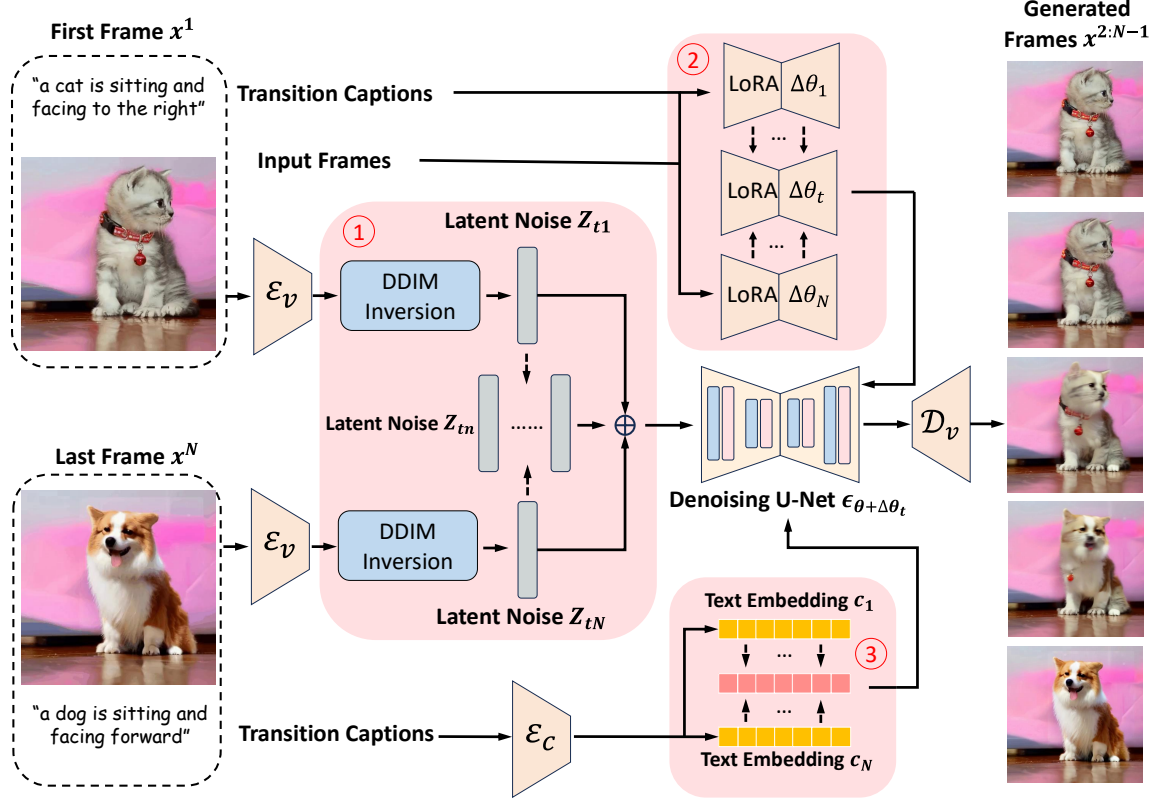
Figure 2. **Inference framework of VTG.** Our interpolation-based initialization features three designs: ① Interpolated Noise Injection, ② LoRA Interpolation, and ③ Frame-aware Text Interpolation. VTG first converts two encoded input frames to latent noises via DDIM inversion. Then, it interpolates between the two latent noises and concatenate intermediate noises along the temporal dimension. To capture meaningful semantics and enable the transition between two conceptually different objects, we employ LoRA interpolation and text interpolation, respectively.

puts. To suppress such transition artifacts, we first train two LoRAs $\Delta\theta_1$, $\Delta\theta_N$ given two input frames x^1 and x^N, optimizing the following objective:

$$\mathcal{L}(\Delta\theta_t) = \mathbb{E}_{\epsilon,t}\big[\|\epsilon - \epsilon_{\theta+\Delta\theta_i}(\sqrt{\bar{\alpha}_t}\,\mathbf{z}_{0n} + \sqrt{1 - \bar{\alpha}_t}\,\epsilon, t, \mathbf{c}_n)\|^2\big], \quad (5)$$

where $\mathbf{z}_{0n}$ is the encoded latent vector of frame n, and $\mathbf{c}_n$ is the text embedding associated with transition caption c^n. Then, we linearly interpolate the two LoRAs to fuse the semantics of two input frames:

$$\Delta\theta = (1 - \lambda_{lora})\Delta\theta_1 + \lambda_{lora}\Delta\theta_N, \quad (6)$$

where λ_{lora} denotes the parameter during lora interpolation.

Typically, video diffusion models [8, 50] employ only one caption as the text condition. This hinders the transition generation between two conceptually different objects, causing abrupt content changes without generating intermediate frames with hybrid meanings. However, for image-to-video diffusion models, the text embedding of the entire frame sequence c is integrated within the denoising U-Net via the cross-attention layer, without explicitly defining per-frame text embedding. A related *still-image* approach [52]

linearly blends the two text embeddings to morph between two static endpoints, whereas our method performs *frame-aware* SLERP across prompt sequences as follows:

$$\mathbf{c}_{\lambda_{text}} = \frac{\sin((1 - \lambda_{text})\phi)}{\sin\phi}\mathbf{c}_1 + \frac{\sin(\lambda_{text}\phi)}{\sin\phi}\mathbf{c}_N, \quad (7)$$

where $\lambda_{text} \in [0, 1]$ serves as the frame-aware coefficient to control the transition sequence; $\mathbf{c}_1$ and $\mathbf{c}_N$ denote the text embeddings of the first and last frames, respectively.

3.3. Bidirectional Motion Prediction

In our experiments, we observed notable quality differences when reversing the order of input frames. This discrepancy arises from two main factors: (1) the model shows a bias toward resembling the initial input frame more closely due to conditional image leakage [59]; (2) existing image-to-video diffusion models [8, 50] are pretrained exclusively for forward motion prediction, while real-world motion is naturally asymmetric, resulting in ambiguity when predicting reverse motion.

Figure 3. **Training framework of VTG.** In Bidirectional Motion Prediction (BMP), the noisy latent is flipped along temporal dimension, and self-attention maps undergo a 180-degree rotation to establish reversed motion-time correlations. Two U-Nets separately predict forward and backward motion, with the backward noise reversed again to fuse with the forward noise, ensuring a consistent motion path for iterative denoising. In Representation Alignment Regularization (RAR), we spatially patchify each frame independently, then aggregate the per-patch alignment loss across temporal dimension.

Inspired by [46], we simultaneously predict forward and backward motions to mitigate the ambiguity issue. Specifically, we rotate the temporal self-attention maps by 180 degrees both horizontally and vertically, causing a reversal of attention relationships as illustrated in Figure 3. We also reverse the original forward trajectory of video $\{x^1, x^2, \dots, x^N\}$ to $\{x^N, x^{N-1}, \dots, x^1\}$, given the two input frames x^1 and x^N. Then, we obtain the reversed noisy latent $\mathbf{z}_{t'}$ by flipping $\mathbf{z}_t$ along the temporal dimension. After that, $\mathbf{z}_{t'}$ is fed into the 3D U-Net for backward motion prediction.

During training, we employ a lightweight fine-tuning strategy that only updates the parameters of value and output matrices (i.e., $\theta_{w,o}$) in the temporal attention layers. The objective $\mathcal{L}_{BMP}$ is formulated by taking the L2 norm between the predicted noise and ground-truth reversed noise:

$$\mathcal{L}_{BMP} = \| \, \mathrm{flip}(\epsilon_t) - \epsilon_{\theta_{w,o}}(\mathbf{z}_{t'}, \mathbf{c}, t, A'_{i,j}) \, \|_2^2 \,. \quad (8)$$

The predicted backward noise is reversed again to fuse with forward noise. Specifically, we simply employ linear interpolation to ensure forward-backward consistency:

$$\epsilon_t = (1 - \lambda_{BMP})\epsilon_{t,i} + \lambda_{BMP}\epsilon'_{t,N-i}, \quad (9)$$

where $\lambda_{BMP} = 0.5$ denotes the weighting factor.

3.4. Representation Alignment Regularization

Though previous techniques effectively mitigate abrupt content changes and motion-reversal ambiguity, our approach remains susceptible to generating blurry and low-fidelity transitions, especially when input frames contain fine-grained textures (e.g., bicycle spokes or fabric weave). Recent work [53] shows that diffusion latents inherently lack high-frequency semantics compared to self-supervised representations like DINOv2's [32]. Such mismatch tends to accumulate over video frames and appears as blur.

To overcome this, we propose to distill DINOv2 features back into the denoising trajectory of video diffusion. As illustrated in Figure 3, the video latent is first patchified into N sequences of T tokens along the temporal dimension. For a N-frame video, the latent vector $\mathbf{z}_n$ of n-th frame has the shape of $C \times h \times w$. Given a pre-defined patch size s, the sequence length T should be: $T = (h/s) \times (w/s)$. Then, we project the per-frame latent representation $\mathbf{z}_n$ of video diffusion via a multilayer perceptron (MLP) to align with DINOv2 representation $\mathbf{y}_* = \mathcal{E}_d(\mathbf{x}_*) \in \mathbb{R}^{P \times m}$, where $\mathbf{y}_*$ denotes the encoder output given a clean video frame $\mathbf{x}_*$; P and m respectively denotes the number of patches and the embedding dimension of DINOv2 encoder $\mathcal{E}_d$. Let $y_\phi(\mathbf{h}_t)$ be the projected representation of intermediate diffusion latent feature $\mathbf{h}_t$, where ϕ denotes the trainable parameters of the MLP. The regularization term $\mathcal{L}_{RAR}$ is computed by

16985

aggregating patch-wise similarities across all video frames:

$$\mathcal{L}_{\text{RAR}} = -\sum_{n=1}^{N} \mathbb{E}_{t,\mathbf{x}_*,\epsilon_t} \left[\frac{1}{P} \sum_{p=1}^{P} \text{sim}\big(\mathbf{y}_*^{[p]}, y_\phi(\mathbf{h}_t)^{[p]}\big) \right]. \quad (10)$$

During inference, the external encoder $\mathcal{E}_d$ and MLP projector y_ϕ are discarded.

4. Experiments

4.1. Experimental Setup

Training Details. We leverage a small collection (i.e., 150) of high-quality in-house videos as our training data. These video samples encompass various motion (e.g., people running, airplanes taxiing) and appearance changes (e.g., a man transitioning from youth to old age). The learnable parameters of the backward U-Net are initialized to zero, while the other available weights are initialized from the pre-trained checkpoint of DynamiCrafter [50]. We utilize the AdamW optimizer with learning rate of $1 \times e^{-5}$. VTG is trained on 4 NVIDIA A100-80GB GPUs, taking $\sim 20K$ iterations with batch size of 2.

Inference Details. In our experiments, we use Stable Diffusion v2.1-base [36] as our interpolation backbone. When training LoRAs, we set the LoRA rank to 16 and train them for 200 steps using AdamW optimizer [28] with a learning rate of 2×10^{-4}. This requires only $\sim 85s$ with 1 NVIDIA A100-80GB GPU. During inversion and sampling, we use the DDIM sampler [42] and set total time steps to the default value (i.e., 50) of the selected image-to-video models. we apply text classifier-free guidance [17] for all models with the same negative prompt "Distorted, discontinuous, Ugly, blurry, low resolution, motionless, static, disfigured, disconnected limbs, Ugly faces, incomplete arms" across all transition generations.

Comparison Methods. We compare VTG with following methods: DiffMorpher [56], TVG [57], SEINE [8], DynamiCrafter [50], Generative Inbetweening [46], Text Embedding Interpolation (TEI) [45], Denoising Interpolation (DI) [16], and Attention Interpolation (AID) [16].

Evaluation Benchmarks. We employ several public datasets and our self-curated TransitBench to comprehensively evaluate the quality of the generated transition videos. Specifically, we leverage MorphBench [56] and TC-Bench [13] to evaluate object morphing and motion prediction. Concept blending and scene transition are rarely explored in the realm of generative models. Hence, there is a lack of specific evaluation benchmarks for these two tasks. Previously, [16] employs CIFAR-10 [24] and LAION-Aesthetics [39] as their evaluation benchmarks to test concept blending. While there are sufficient number of

samples in both datasets, CIFAR-10 only contains concepts from 10 classes, and LAION-Aesthetics consists entirely of synthetic images and lacks real-world samples. To this end, we present *TransitBench*, the first benchmark dataset for collectively assessing concept blending transitions of two distinct conceptual objects and scene transitions between two relevant scenarios. We collected 200 pairs of pictures (each pair forms the first and the last frames of one transition generation sample) of diverse content and styles, and evenly divide them into two categories: 1) *concept-blending* cases, and 2) *scene-transition* cases, both of which are obtained from web resources. We hope TransitBench can promote future studies on general transition generation.

4.2. Qualitative Results

To demonstrate the superiority of our method, we provide a visual comparison of VTG against five state-of-the-art baselines. Specifically, we compare our approach with Diff-Morpher, Generative Inbetweening, TVG, SEINE, DynamiCrafter across four transition tasks for uniformity. Figure 4 shows the generation results by different methods. For object morphing, the compared methods exhibit low-fidelity (e.g., DiffMorpher, DynamiCrafter), oversaturation effect (e.g., TVG), and inaccurate semantics (e.g., Generative Inbetweening, SEINE). In contrast, VTG presents a natural transition, preserving the content of the first and last frames well. For motion prediction, Generative Inbetweening and SEINE generate inferior transitions as observed by their blurry intermediate frames. TVG yields unnatural transitions as it generates a front-facing image of a person, while the person's face is actually turned away. DiffMorpher shows low-fidelity transitions, and its generated video looks like a sequence of intermittent images without concerning the consistency between background and foreground objects. This is largely due to the missing temporal modeling in DiffMorpher's image diffusion backbone. VTG achieves comparable or even better results with DynamiCrafter, validating that it can effectively preserve the appearance and motion priors in diffusion models. For concept blending, most video diffusion baselines present abrupt content changes, while other baselines show blurry (e.g., DiffMorpher) and ambiguous (e.g., TVG) intermediate frames. In contrast, VTG yields smooth transition between the lion and the truck with rational intermediate results (e.g., a lion-colored and lion-sized truck). For scene transition between two relevant scene images, existing methods suffer from abrupt content changes or low similarity (e.g., TVG), while VTG produces natural scene-level transition with pleasing effect.

4.3. Quantitative Results

Evaluation Metrics. The objective of *versatile transition generation* is to generate a sequence of $(N - 2)$ intermedi-
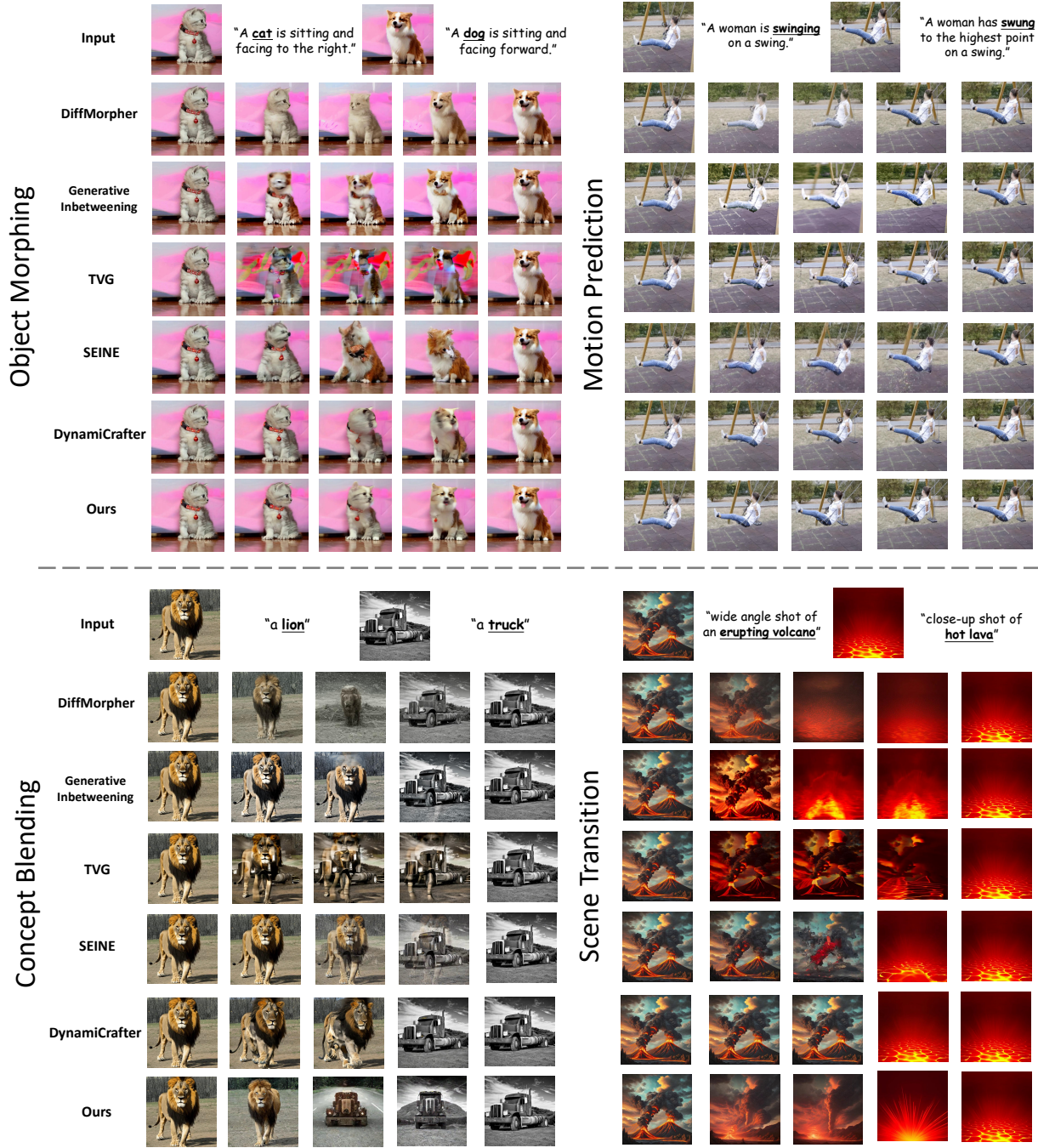
Figure 4. **Qualitative comparisons on four transition tasks.** For each task, the displayed images correspond to the 1st, 4th, 8th, 12th, and 16th video frames, respectively. Please refer to our project page for more video examples, which offers a more straightforward and comprehensive qualitative comparisons.

ate video frames, denoted as $x^{2:N-1}$, between two source frames x^1 and x^N, with two captions describing the two frames. High-quality transition videos should simultaneously satisfy four criteria: semantic fidelity to the input endpoints, visual quality, temporal coherence, and perceptual

smoothness. To ensure fair comparisons, we performed extensive quantitative evaluations using experimental settings consistent with those prior methods as in their original paper. Specifically, we adopt the following evaluation metrics: (1) FID (↓) and PPL (↓) follow DiffMorpher [56]. FID com-

pares the first and last frames against the generated transition frames in the Inception feature space; PPL measures the average LPIPS fluctuation along an interpolation path, where lower values imply fewer perceptual jumps. (2) TCR (↑) and TC-Score (↑) are taken from the training-free TVG framework [57]. TCR is the fraction of video frames whose CLIP embedding stays within a pre-set similarity margin to either endpoint; TC-Score complements it with a mean cosine-similarity, offering a softer view of frame-level consistency. (3) Smoothness (↑) is adapted from AID [16]. For every adjacent frame pair, we compute LPIPS and the Gini coefficient of these distances, reflecting how even the step-to-step changes are. We report its inverse so that a higher score denotes smoother transitions.

Method	Metamorphosis		Animation	
	FID (↓)	PPL (↓)	FID (↓)	PPL (↓)
DiffMorpher [56]	<u>70.49</u>	**18.19**	<u>43.15</u>	**5.14**
TVG [57]	86.92	35.18	42.99	12.46
SEINE [8]	82.03	47.72	48.25	16.26
DynamiCrafter [50]	87.32	42.09	43.31	<u>11.16</u>
VTG (Ours)	**67.39**	<u>22.80</u>	**39.16**	**5.14**

Table 1. **Quantitative results on MorphBench.** The best results are in **bold**; second-best are <u>underlined</u>.

Method	Attribute		Object		Background	
	TCR (↑)	TC-Score (↑)	TCR (↑)	TC-Score (↑)	TCR (↑)	TC-Score (↑)
DiffMorpher [56]	41.82	0.844	19.57	0.765	50.00	0.819
SEINE [8]	17.86	0.720	10.48	0.654	7.96	0.742
DynamiCrafter [50]	16.55	0.745	13.91	0.707	25.56	0.795
TVG [57]	41.82	0.877	30.44	0.822	38.89	0.864
VTG (Ours)	**42.78**	**0.893**	**33.46**	**0.849**	**50.00**	**0.883**

Table 2. **Quantitative results on TC-Bench.** The best results are in **bold**. Best viewed when zoomed in.

Dataset	TEI	DI	AID-O	AID-I	Ours
CIFAR-10	0.7531	0.7564	0.7831	0.7861	**0.7932**
LAION-Aesthetics	0.7424	0.7511	0.7643	0.8152	**0.8215**

Table 3. **Smoothness (↑) evaluation for Concept Blending.** The best results are in **bold**.

Quantitative Results. As seen in Table 1 and Table 2, VTG outperforms existing methods with their original experimental setting, demonstrating its capability to generate high-quality, temporally coherent, and semantically consistent video transitions. It is worth noting that the only exception is *Metamorphosis* subset, where VTG achieves the second-best PPL (22.80). Nevertheless, VTG still demonstrates outstanding overall performance, achieving state-of-the-art FID and comparable PPL.

According to AID [16], a well-interpolated concept-blending sequence should exhibit a gradual and smooth

transition. To offer a more comprehensive evaluation, we conducted experiments using the same benchmarks and metric as in AID. Please refer to Table 3 for details.
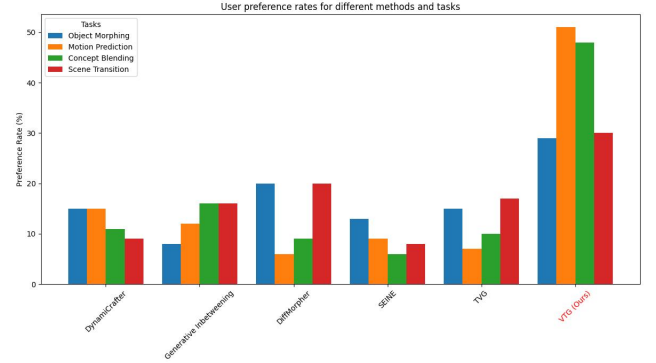

Figure 5. **The bar plot of user preference rates (%) of different methods on four transition tasks.** Best viewed when zoomed in.

User Study. To subjectively evaluate our approach, we further conduct a user study. We asked Amazon Mechanical Turk (AMT) workers to choose the best generated transition video from a set of five candidates. Given the two input frames and corresponding transition captions, AMT workers evaluate the videos based on three questions: 1) *Which video is the most semantic-coherent?* 2) *Which video is the most temporally coherent?* 3) *Which video has the highest fidelity?* To make a fair evaluation, we include static and obviously incorrect videos to identify random clicking. Each test example is assigned to 10 different workers for evaluation. We collected 150 valid responses for each transition task after filtering out those responses from random clicking workers. The results are shown in Figure 5, which illustrates that VTG obtains the highest preference rate with its superior quality of the generated transition videos compared to other baselines.

5. Conclusion

In this paper, we formulate the problem of *versatile transition generation* that encompasses four representative transition tasks, namely, object morphing, motion prediction, concept blending, and scene transition. Leveraging a pre-trained image-to-video diffusion model, we design **VTG**, a unified **V**ersatile **T**ransition video **G**enerator that can tackle the four tasks within a single framework. Both qualitative and quantitative experiments show that VTG achieves superior performance in generating semantically relevant, temporally coherent, and visually pleasing transition video frames, given a pair of text prompts and two input frames. With these advantages, VTG can be readily applied to real-world content creation scenarios, offering an efficient tool for generating high-quality transition in media production.

Acknowledgements

References

[1] Alyaa Qusay Aloraibi. Image morphing techniques: A review. *Technium: Romanian Journal of Applied Sciences and Technology*, 2023. 2

[2] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. *Seminal Graphics Papers: Pushing the Boundaries*, 1992. 2

[3] Bhumika G Bhatt. Comparative study of triangulation based and feature based image morphing. *Signal & Image Processing*, 2011. 2

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

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[6] Erik Bodin, Alexandru I Stere, Dragos D Margineantu, Carl Henrik Ek, and Henry Moss. Linear combinations of latents in generative models: subspaces and beyond. In *ICLR*, 2025. 3

[7] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 1

[8] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2023. 2, 3, 4, 6, 8

[9] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *TOG*, 2012. 2

[10] Yitong Deng, Winnie Lin, Lingxiao Li, Dmitriy Smirnov, Ryan Burgert, Ning Yu, Vincent Dedun, and Mohammad H Taghavi. Infinite-resolution integral noise warping for diffusion models. In *ICLR*, 2025. 3

[11] Jiong Dong, Kaoru Ota, and Mianxiong Dong. Video frame interpolation: A comprehensive survey. *TOMM*, 2023. 2

[12] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *ECCV*, 2024. 3

[13] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. In *ACL*, 2025. 6

[14] Noa Fish, Richard Zhang, Lilach Perry, Daniel Cohen-Or, Eli Shechtman, and Connelly Barnes. Image morphing with perceptual constraints and stn alignment. In *Computer Graphics Forum*, 2020. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[16] Qiyuan He, Jinghao Wang, Ziwei Liu, and Angela Yao. Aid: Attention interpolation of text-to-image diffusion. In *NeurIPS*, 2024. 2, 6, 8

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 3

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[21] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, 2022. 3

[22] Siddhant Jain, Daniel Watson, Eric Tabellion, Aleksander Hołyński, Ben Poole, and Janne Kontkanen. Video interpolation with diffusion models. In *CVPR*, 2024. 2, 3

[23] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *CVPR*, 2022. 3

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009. 6

[25] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *CVPR*, 2023. 2, 3

[26] Jing Liao, Rodolfo S Lima, Diego Nehab, Hugues Hoppe, Pedro V Sander, and Jinhui Yu. Automating image morphing using structural similarity on a halfway domain. *TOG*, 2014. 2

[27] Chunxu Liu, Guozhen Zhang, Rui Zhao, and Limin Wang. Sparse global matching for video frame interpolation with large motion. In *CVPR*, 2024. 2, 3

[28] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 2018. 6

[29] Michael I. Miller and Laurent Younes. Group actions, homeomorphisms, and matching: A general framework. *IJCV*, 2001. 2

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and

Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2021. 1

[31] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, 2020. 2

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 5

[33] Marko Rajković. *Geodesics, Splines, and Embeddings in Spaces of Images*. PhD thesis, Universitäts-und Landesbibliothek Bonn, 2023. 2

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[35] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *ECCV*, 2022. 2, 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 6

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 6

[40] Eli Shechtman, Alex Rav-Acha, Michal Irani, and Steve Seitz. Regenerative morphing. In *CVPR*, 2010. 2

[41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 1

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 3, 6

[43] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 1

[44] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, 2024. 1

[45] Clinton Wang and Polina Golland. Interpolating between images with diffusion models. In *ICMLW*, 2023. 6

[46] Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steven M Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In *ICLR*, 2025. 3, 5, 6

[47] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, 2024. 1

[48] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, 2024. 1

[49] George Wolberg. Image morphing: a survey. *The visual computer*, 1998. 2

[50] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, 2024. 2, 3, 4, 6, 8

[51] Cheng Yang, Lijing Liang, and Zhixun Su. Real-world denoising via diffusion model. *arXiv preprint arXiv:2305.04457*, 2023. 2

[52] Zhaoyuan Yang, Zhengyang Yu, Zhiwei Xu, Jaskirat Singh, Jing Zhang, Dylan Campbell, Peter Tu, and Richard Hartley. Impus: Image morphing with perceptually-uniform sampling using diffusion models. In *ICLR*, 2024. 2, 4

[53] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025. 5

[54] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *CVPR*, 2024. 2, 3

[55] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *CVPR*, 2023. 2

[56] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Xingang Pan, and Bo Dai. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *CVPR*, 2024. 2, 3, 6, 7, 8

[57] Rui Zhang, Yaosen Chen, Yuegen Liu, Wei Wang, Xuming Wen, and Hongxia Wang. Tvg: A training-free transition video generation method with diffusion models. *TCSVT*, 2024. 3, 6, 8

[58] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 1

[59] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. In *NeurIPS*, 2024. 4

[60] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 1

[61] Bhushan Zope and Soniya B Zope. A survey of morphing techniques. *IJAEMS*, 2017. 2