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.

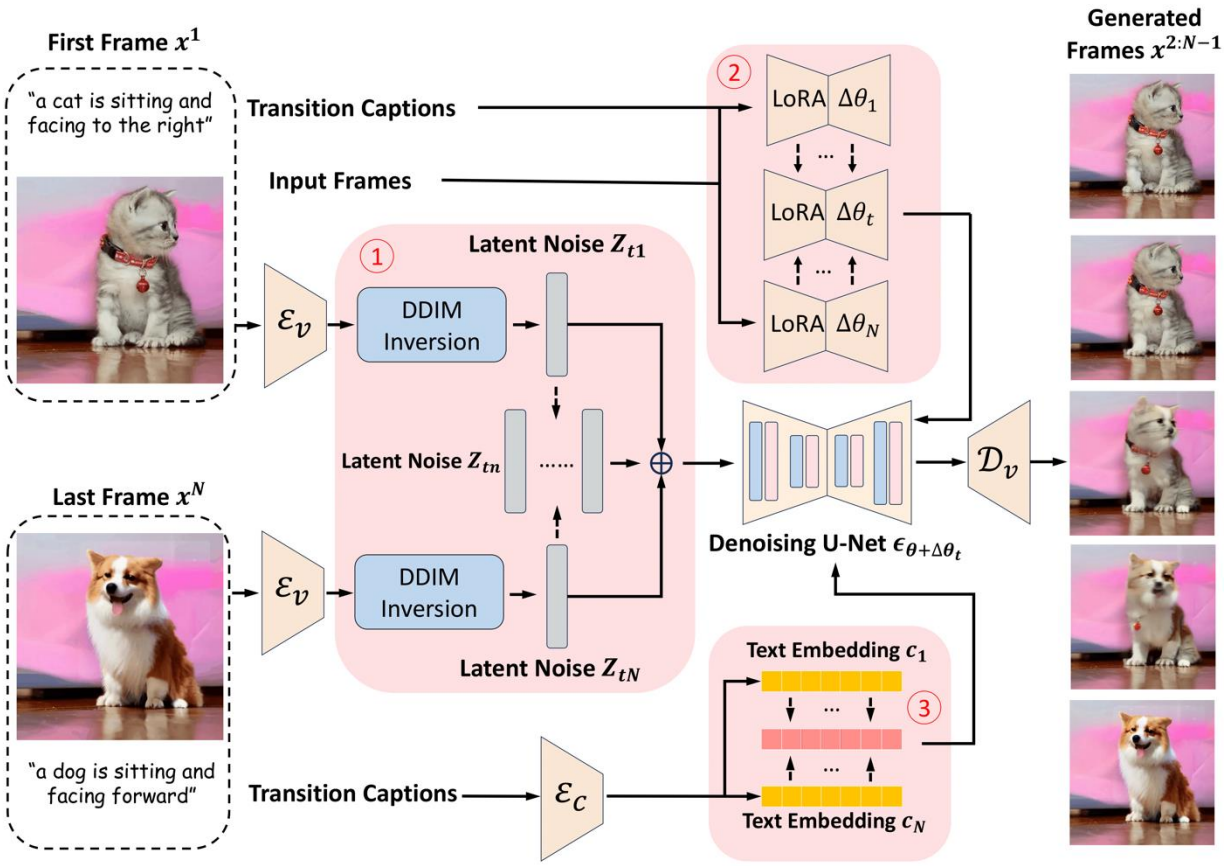
[*]Work was done while interning at ByteDance [†]Corresponding Author

Background & Motivation

Video editors often need smooth and high-fidelity transitions when connecting disparate scenes. Existing diffusion models excel at image synthesis but struggle to interpolate semantically different endpoints with temporal coherence.

We seek to unify four tasks – (a) object morphing, (b) motion prediction, (c) concept blending, and (d) scene transition – under a single image-to-video diffusion framework.

Interpolation-based Initialization



During inference, we do …

Noise Interpolation

Spherical linear interpolation between the latent noises of the first and last frames suppresses flicker and preserves structure.

LoRA Interpolation

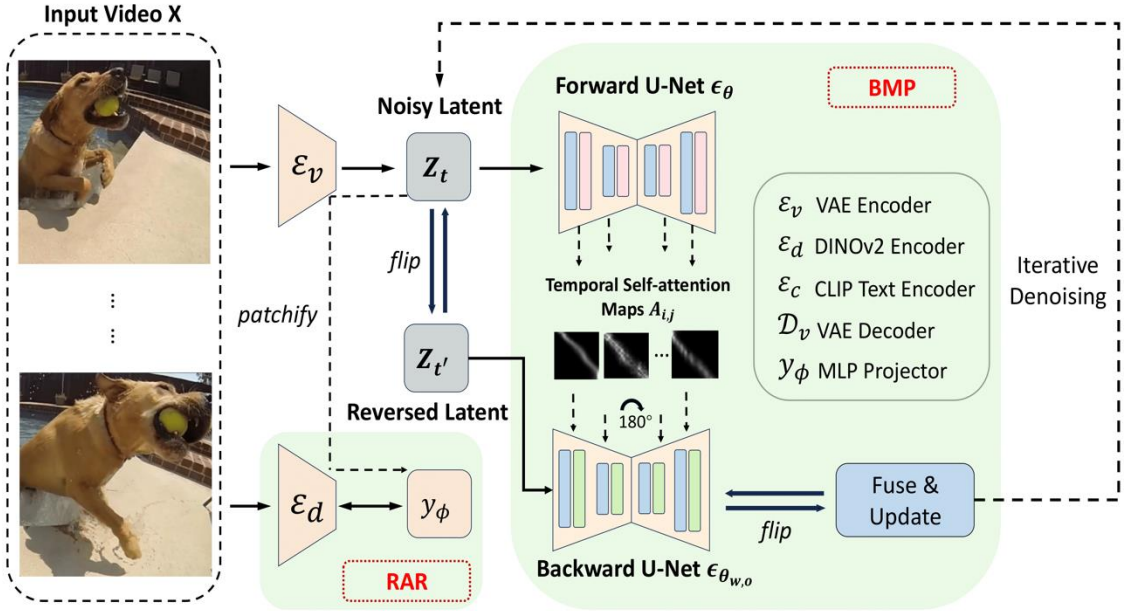
Two LoRA-integrated U-Nets encode object semantics from both endpoints.

Text Interpolation

Frame-aware interpolation of text embeddings yield intermediate frames with hybrid textual meanings.

Bidirectional Motion Prediction & Representation Alignment Regularization

During training, we do …



Bidirectional Motion Prediction

Flip the latent sequence, rotate self-attention maps by 180° and merge backward predictions to ensure consistent motion trajectories.

Representation Alignment Regularization

Patchify each frame, compute per-patch alignment losses and aggregate across time to encourage visual consistency and fidelity.

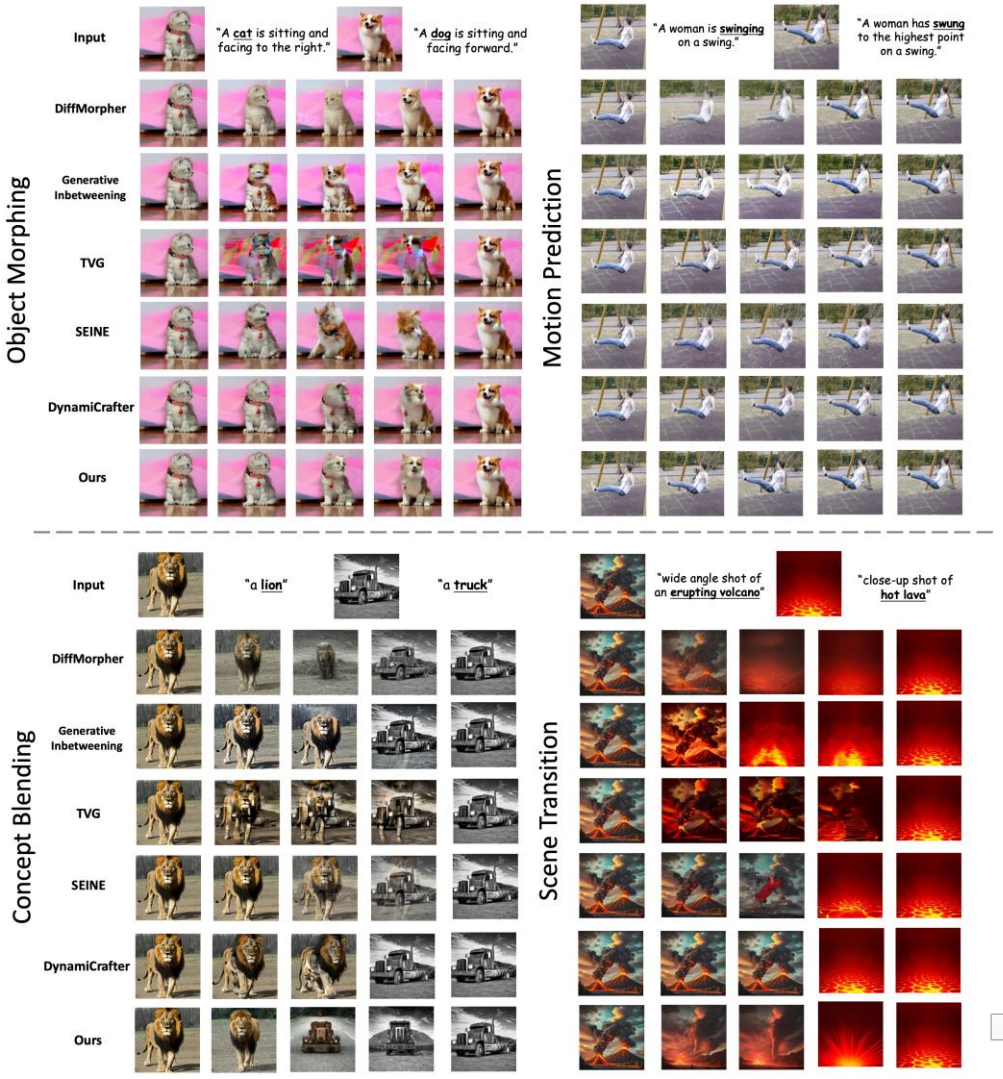
Qualitative Comparisons

In the **object morphing**, our approach produces naturally evolving transformations that faithfully preserve the structure and semantics of the endpoint frames, while avoiding oversaturation or semantic drift—demonstrating high-fidelity and stable morph transitions.

In the **motion prediction**, our method yields intermediate frames that are sharp and coherent, maintaining consistent motion, and seamlessly integrating foreground and background elements to achieve a smooth dynamic progression.

In the **concept blending**, our framework generates intermediate representations that smoothly and meaningfully mediate between two distinct concepts, enabling a gradual and semantically coherent fusion without abrupt shifts.

In the **scene transition**, our approach constructs visually harmonious sequences between two related scenes, delivering content continuity, compositional consistency, and a pleasing, natural transition.



Quantitative Comparisons

Across all evaluated tasks, our approach consistently demonstrates superior or comparable performance to existing methods under both perceptual and temporal metrics.

In the **Table 1**, our approach achieves the lowest FID and one of the lowest PPL values, indicating its ability to generate high-quality and temporally coherent transition sequences. Notably, it attains state-of-the-art FID while maintaining competitive perceptual smoothness, reflecting its effectiveness in preserving both content fidelity and motion dynamics during transitions.

In the **Table 2**, our method yields the highest TCR and TC-Scores across all categories. These metrics, derived from cosine similarity and temporal consistency measures, highlight the method's strength in maintaining semantic coherence and smooth frame-to-frame evolution throughout the transition process.

In the **Table 3**, our framework achieves the highest smoothness scores across different datasets, surpassing baseline methods in generating transitions that are visually continuous and conceptually well-integrated. This confirms its ability to handle more abstract blending scenarios with balanced temporal smoothness and semantic plausibility.

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>.

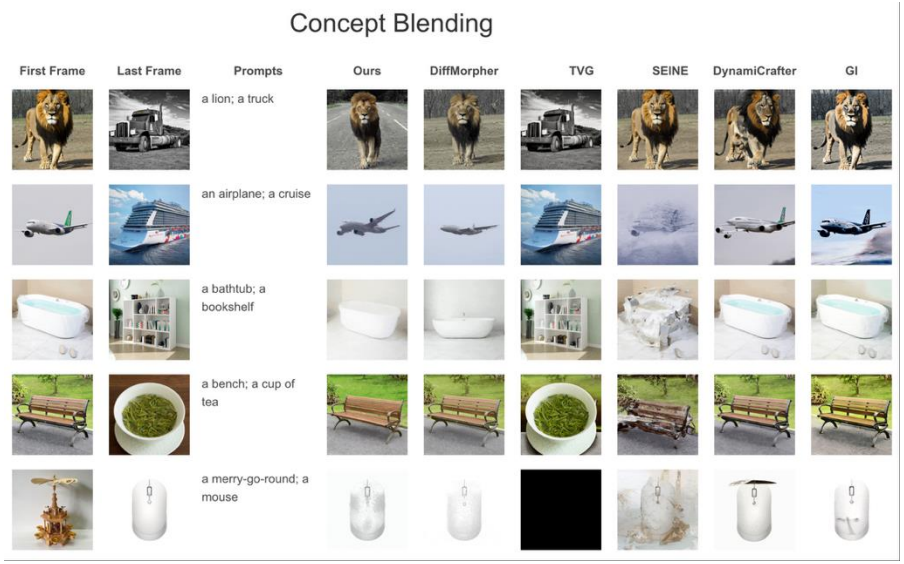
Method	*Attribute*		*Object*		*Background*	
	TCR (↑)	TC-Score (↑)	TCR (↑)	TC-Score (↑)	TCR (↑)	TC-Score (↑)
DiffMorpher [56]	41.82	0.844	19.57	0.765	**50.00**	0.819
SEINE [8]	17.86	0.720	10.48	0.654	7.96	0.742
DynamiCrafter [50]	16.55	0.745	13.91	0.707	25.56	0.795
TVG [57]	41.82	0.877	30.44	0.822	38.89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

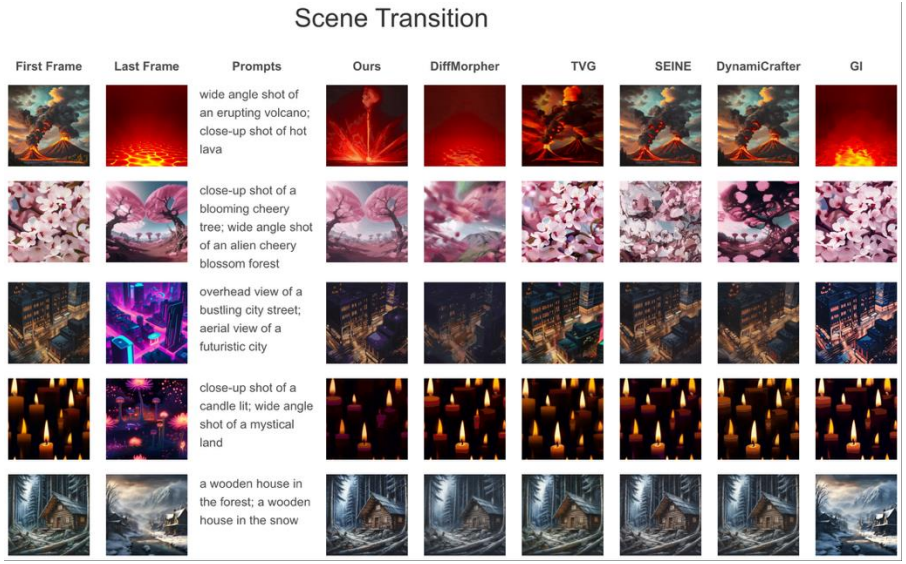
Table 2. **Quantitative results on TC-Bench.** The best results are in **bold**. Best viewed when zoomed in.

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**.

TransitBench





We introduce **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.

Conclusion

- Unified four transition tasks under a single diffusion-based framework
- Three interpolation-based initialization modules for coherent transitions
- Bidirectional Motion Prediction & Representation Alignment Regularization for better smoothness and fidelity
- Introduced TransitBench to benchmark Concept Blending and Scene Transition

Future Directions

- Longer and more complex transitions
- Real-time insertion for streaming video generation

Thanks for watching this video!