# Bidirectional Attention-Gated Motion Injection for Frame Interpolation

Ziyue Zeng
*Graduate School of FSE*
*Waseda University*
Tokyo, Japan
zengziyue@fuji.waseda.jp

Yui Tatsumi
*Graduate School of FSE*
*Waseda University*
Tokyo, Japan
yui.t@fuji.waseda.jp

Hiroshi Watanabe
*Graduate School of FSE*
*Waseda University*
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—We propose Bi-AGMI (Bidirectional Attention-Gated Motion Injection), a lightweight and efficient framework for keyframe interpolation based on diffusion models. Bi-AGMI introduces a dual-path denoising process that sequentially connects forward and backward sampling trajectories via latent flipping, enabling temporally bounded generation from two keyframes. To enhance consistency between these two trajectories, we design a novel attention-gated fusion mechanism that dynamically injects and blends forward-path attention features into the backward UNet using a learnable gating module. This design improves temporal coherence, mitigates motion ambiguity, and eliminates the need for repeated re-noising. Experiments on DAVIS and Pexels datasets demonstrate that our method achieves competitive visual quality and inference efficiency compared to recent diffusion-based baselines, while requiring significantly fewer sampling steps. By enabling stable interpolation over large temporal gaps, Bi-AGMI expands the practical usability of diffusion models for long-range video completion.**

***Index Terms*—Diffusion models, frame interpolation.**

## I. Introduction

Frame interpolation aims to synthesize a sequence of intermediate frames between two input keyframes, enabling the reconstruction of smooth and temporally consistent video. It plays a vital role in applications such as frame rate upsampling, slow-motion generation, and video restoration. Recently, diffusion-based generative models have shown strong potential for this task by leveraging rich motion priors learned from large-scale datasets [2]–[4].

While these models produce high-fidelity frames, challenges remain in maintaining temporal coherence and efficiency—particularly for long-range interpolation where the input frames are far apart or contain ambiguous motion. Several recent methods have extended image-to-video diffusion models to support dual-keyframe conditioning [5]–[7], typically by designing new sampling or feature fusion strategies. However, these approaches either lack direct cross-trajectory communication or rely on fine-tuning and multi-pass inference, limiting their efficiency and robustness.

In this work, we propose **Bi-AGMI** (Bidirectional Attention-Gated Motion Injection), a lightweight and training-free framework for keyframe interpolation. Unlike prior bidirectional methods that either lack interaction between forward and backward paths [5], or rely on weight sharing and repeated re denoising [6], Bi-AGMI introduces a gated attention fusion mechanism that directly injects forward-path temporal features into the backward UNet through learnable gates.

This design brings three key advantages: (1) *Cross-path interaction without model fine-tuning* — attention features are fused at inference-time via a parameter-efficient gate without modifying pretrained weights; (2) *Layer-wise control and flexibility* — our injection is configurable per attention layer, allowing for minimal and effective intervention in the backbone; (3) *Efficient single-pass inference* — the method avoids iterative re-noising and completes generation in a single forward-backward denoising cycle.

Extensive experiments on DAVIS and Pexels datasets demonstrate that Bi-AGMI achieves competitive visual quality and motion consistency with significantly reduced inference cost. Notably, it performs robustly in long-interval interpolation scenarios, enhancing the practical usability of diffusion models for video completion tasks.

### A. Related Work

**Explorative Inbetweening.** Feng et al. [1] first introduced the concept of *bounded generation*, which frames keyframe interpolation as generating a video sequence bounded by given start and end frames. They propose Time Reversal Fusion (TRF), a training-free sampling strategy that fuses forward and backward denoising paths in Stable Video Diffusion to synthesize smooth transitions. This work highlights the potential of leveraging pretrained diffusion models for controllable video generation without fine-tuning, and has since inspired a series of dual-path interpolation methods.

**Dual-Path Diffusion for Interpolation.** Several recent approaches have extended this idea by exploring different fusion strategies between forward and backward sampling paths. ViBiDSampler [5] performs sequential denoising from both directions with a single re-noising step in between, improving consistency without modifying the model. Generative Inbetweening [6] (SVD-Kframe) shares temporal self-attention maps across dual UNets to encourage symmetry but requires fine-tuning and multiple re-denoising passes. FCVG [7] further introduces explicit visual conditions between frames to guide the diffusion process, enhancing alignment but adding structural complexity. While these methods improve motion coherence, they still suffer from limitations such as lack of

979-8-3315-2415-9/25/$31.00 ©2025 IEEE

direct interaction between trajectories, high inference cost, or reliance on additional supervision.

To address these issues, we propose Bi-AGMI, which enables direct cross-path information flow by injecting forward-path attention features into the backward UNet via a learnable gating module.

## II. Bidirectional Attention-Gated Motion Injection

### A. Dual-Path Inference with Latent Flip

The core of our framework is a bidirectional denoising process that sequentially generates forward and backward trajectories. Given a random initial latent $\mathbf{z}_t \sim \mathcal{N}(0, I)$, we first perform standard forward denoising conditioned on the start frame $I_{\text{start}}$ using the original UNet (Pre-trained stable video diffusion). The noise prediction $\hat{\epsilon}_f$ is estimated and passed into a diffusion scheduler (e.g., Euler method) to obtain the partially denoised latent:

$$\mathbf{z}_{t-1} = \text{Step}(\mathbf{z}_t, \hat{\epsilon}_f). \tag{1}$$

To initiate backward sampling, we re-noise the intermediate latent $\mathbf{z}_{t-1}$ using a newly sampled noise term $\epsilon \sim \mathcal{N}(0, I)$, scaled by the variance gap between adjacent steps. Importantly, a temporal flip operation is applied to reverse the latent sequence:

$$\mathbf{z}'_t = \text{Flip}\left(\mathbf{z}_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \cdot \epsilon\right). \tag{2}$$

The flipped latent $\mathbf{z}'_t$ is then passed into a second UNet, which is conditioned on the end frame $I_{\text{end}}$. A corresponding noise prediction $\hat{\epsilon}_b$ is obtained and used to compute the backward denoised latent:

$$\mathbf{z}'_{t-1} = \text{Step}(\mathbf{z}'_t, \hat{\epsilon}_b). \tag{3}$$

This dual-path inference scheme enables a one-pass forward-backward generation process bounded by keyframes, and serves as the foundation for our cross-path attention fusion mechanism introduced in the next section.

### B. Bidirectional Attention-Gated Fusion

To facilitate information exchange between the forward and backward denoising paths, we introduce a bidirectional attention-gated fusion module that operates at the attention level within the temporal self-attention layers, as illustrated in Fig. 1. This mechanism serves as the core of Bi-AGMI, enabling directional feature communication without modifying the underlying UNet structure.

During the forward generation phase, we extract temporal attention features $(Q_f, K_f, V_f)$ at selected layers and reverse them along the temporal axis to align with the backward sampling order, denoted as $(Q_{fr}, K_{fr}, V_{fr})$. These features are cached and reused during the backward pass.

In the backward generation phase, we compute two forms of cross-attention using the reversed forward features and the current backward features. The first attends to the backward context using the forward query:

$$A_1 = \text{Attention}(Q_{fr}, K_b, V_b), \tag{4}$$

while the second attends to the forward context using the backward query:

$$A_2 = \text{Attention}(Q_b, K_{fr}, V_{fr}). \tag{5}$$

The final output is a convex combination of these two branches, controlled by a learnable sigmoid gate $\alpha$:

$$\text{Output} = \sigma(\alpha) \cdot A_1 + (1 - \sigma(\alpha)) \cdot A_2, \tag{6}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function. This gated fusion allows the model to adaptively balance contributions from both directions based on motion ambiguity at each layer, enhancing temporal coherence while preserving the efficiency of single-pass inference.

### C. Lightweight Training Strategy

To ensure training efficiency and compatibility with existing diffusion models, we adopt a lightweight optimization strategy. Specifically, we freeze all parameters of the forward and backward UNets and train only the fusion gate parameters $\alpha$ introduced in the attention-level injection module. This selective fine-tuning scheme significantly reduces the number of trainable parameters and avoids disrupting the pretrained generative prior.

The training is performed under mixed-precision (fp16) settings to further accelerate convergence and reduce memory consumption. Since our method builds on top of pretrained backbones and only introduces a small gating layer, it can be trained with minimal computational cost and converges rapidly, even on limited hardware.

## III. Experiments

### A. Experimental Settings and Implementation Details

We evaluate Bi-AGMI on the DAVIS and a custom Pexels dataset, covering diverse scenes and motion patterns. All experiments are conducted using Stable Video Diffusion (SVD-XT) as the backbone, generating 25 intermediate frames at a resolution of $1024 \times 576$, conditioned on the first keyframe.

The underlying UNet contains temporal attention blocks across downsampling, mid, and upsampling stages. We inject forward-path attention features into eight selected layers, choosing the first temporal transformer block (i.e., `attentions.0`) at each stage while skipping secondary blocks to reduce redundancy. Attention dimensions are matched to each stage (320, 640, 1280). Forward attention $(Q_f, K_f, V_f)$ is cached via AttnProcessor hooks and reused in the backward pass for gated fusion.

The injection mechanism is layer-wise configurable. We find that using only the four symmetric layers in the downsampling and upsampling paths offers the best trade-off between quality and stability, and use this as the default in all experiments.

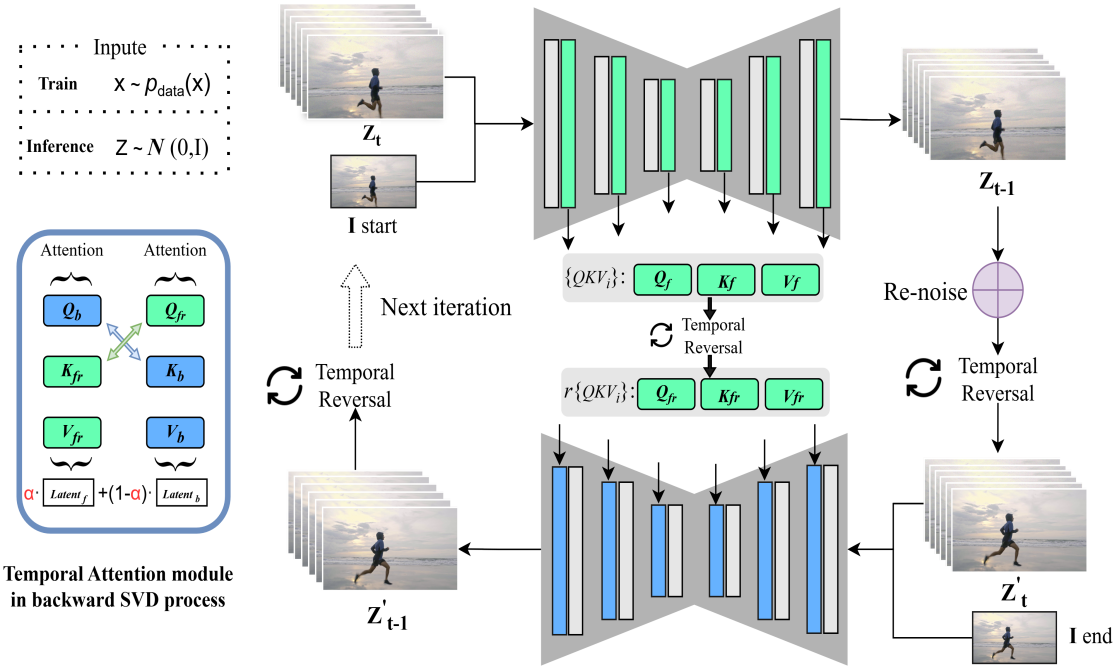


Fig. 1. The Bi-AGMI pipeline, highlighting the iterative sampling loop (right), the inputs during training and inference (upper left), and the temporal attention handling in the backward SVD (lower left).

We adopt an Euler scheduler with 25 denoising steps. Only the gating parameters $\alpha$ are optimized, while all UNet weights remain frozen.

### B. Comparative Studies

We first present a challenging interpolation example for comparison. As shown in Fig 2, both SVD-Kframe and ViBiD methods exhibit structural collapse when the motorcycle passes in front of the background car. In contrast, our method successfully maintains temporal coherence, with the human subject smoothly traversing past the parked car without introducing noticeable artifacts.

Table I compares generation time across methods. Bi-AGMI achieves a substantial speed-up over SVD-Kframe and remains competitive with ViBiD. This efficiency gain stems from our one-pass sampling strategy, which avoids repeated noise injection and denoising cycles. In contrast, SVD-Kframe requires multiple rounds of re-noising to reconcile inconsistencies between forward and backward paths, significantly increasing inference time.

TABLE I
GENERATION TIME COMPARISON

| **1024×576** | **Generation Method** | | |
|---|---|---|---|
| **25 Frames** | SVD-Kframe | ViBiD | Bi-AGMI |
| **Time (s)** | 3049 | 417 | 452 |

### C. Quantitative Evaluation

We compare Bi-AGMI with recent diffusion-based frame interpolation methods, including SVD-Kframe [6] and ViBiDSampler [5], on the DAVIS dataset. Following the evaluation protocols used in both works, we construct 100 keyframe pairs by sampling 25-frame video clips, where the first and last

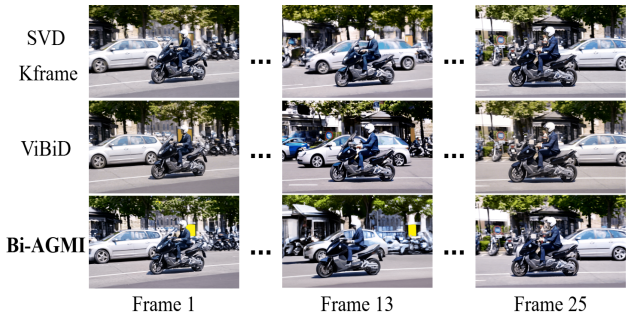


Fig. 2. Example of 25-frame interval video interpolation.

frames serve as input and the remaining 23 frames are reconstructed. All models generate 25-frame sequences conditioned on the first frame, using the same SVD-XT backbone, spatial resolution ($1024 \times 576$), and 25-step Euler sampler. We adopt standard perceptual and distributional metrics—LPIPS, FID, and FVD—to evaluate the fidelity and temporal consistency of the generated results.

Table II reports the perceptual and distributional performance on the DAVIS dataset. Bi-AGMI achieves the best LPIPS score (0.2465), indicating improved perceptual similarity to the ground truth. While SVD-Kframe slightly outperforms in FID and FVD, it requires nearly 7× longer inference time than our method (3049 s vs. 452 s). Compared to ViBiDSampler, Bi-AGMI delivers lower LPIPS, better FVD, and significantly better FID, with only a marginal increase in runtime (452 s vs. 417 s).

These results demonstrate that Bi-AGMI strikes an effective balance between generation quality and efficiency, producing perceptually sharper and temporally consistent results at a

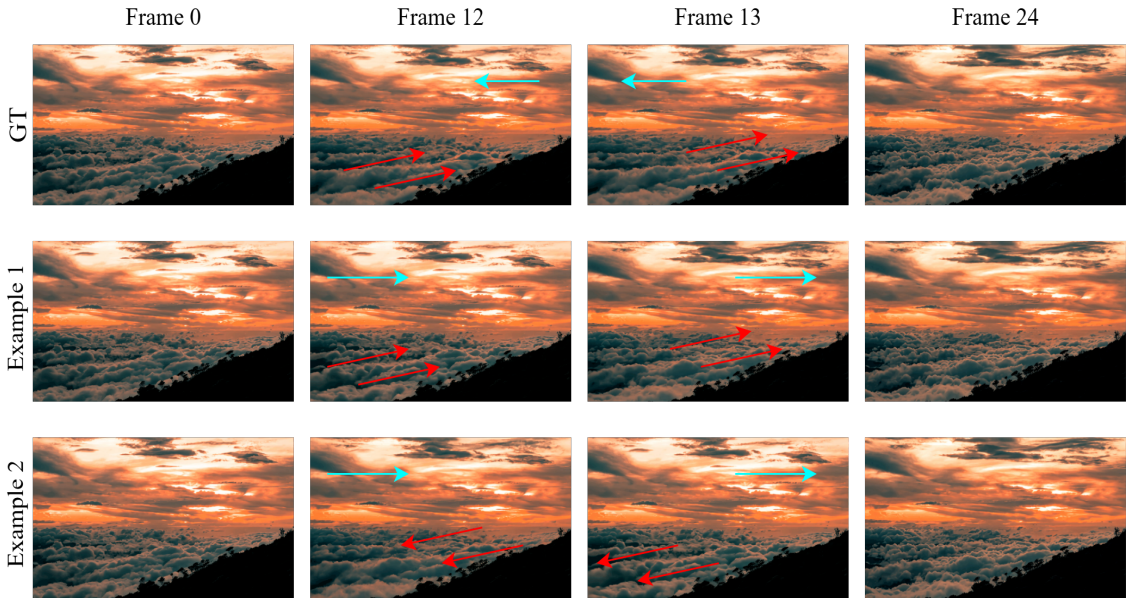


Fig. 3. Example of 25-frame interval video interpolation.

TABLE II
PERCEPTUAL AND DISTRIBUTIONAL METRICS ON DAVIS.

| Method | LPIPS ↓ | FID ↓ | FVD ↓ | Time (s) ↓ |
|---|---|---|---|---|
| SVD-Kframe | 0.2493 | **32.68** | **424.69** | 3049 |
| ViBiDSampler | 0.2574 | 40.72 | 432.21 | **417** |
| **Bi-AGMI (ours)** | **0.2465** | 33.56 | 425.92 | 452 |

TABLE III
ABLATION ON ATTENTION INJECTION DEPTH. WE REPORT FVD ON PEXELS; LOWER IS BETTER.

| Injection Configuration | FVD ↓ |
|---|---|
| No injection | 522.14 |
| Only $8^2$ (mid) | 454.32 |
| $8^2$ + $16^2$ | 439.71 |
| $8^2$ + $16^2$ + $32^2$ | **425.92** |
| Full injection | 428.46 |

fraction of the computational cost required by fully fine-tuned baselines like SVD-Kframe.

### D. Ablation Study

To evaluate the effect of our attention injection strategy, we conduct a controlled ablation by incrementally enabling cross-path attention fusion at different spatial resolutions. All experiments are conducted under the same settings as our quantitative evaluation in Section C, but using the Pexels dataset and report only the FVD score to directly measure temporal consistency.

In our implementation, attention processors are inserted at temporal transformer blocks within the UNet architecture. These injection points correspond to four spatial resolutions, each mapped to specific modules:

$8^2$ (mid_block)
$16^2$ (down_blocks.2 / up_blocks.1)
$32^2$ (down_blocks.1 / up_blocks.2)
$64^2$ (down_blocks.0 / up_blocks.3)

We experiment with four configurations: (1) no injection; (2) mid-only ($8^2$); (3) mid+$16^2$; and (4) full symmetric injection across all levels. Table III presents the ablation results on FVD when enabling attention injection at progressively more spatial layers. We observe that injecting only at the mid-block ($8^2$) already brings significant improvement over the baseline without injection. Further gains are achieved by including the $16^2$ and $32^2$ levels, with the best FVD observed when injecting at three levels ($8^2$–$32^2$). Interestingly, enabling full-layer injection up to $64^2$ slightly degrades performance, suggesting that shallow layers may introduce noise or interfere with deeper temporal modeling.

These results highlight that our method, despite relying on only a few learnable gating parameters, is sensitive to the choice of injection depth. Careful selection of fusion layers is thus critical, and deeper is not always better when working with lightweight controllers.

### E. Limitations

Despite improving temporal smoothness and bidirectional coherence, Bi-AGMI still inherits a fundamental limitation from diffusion-based generation pipelines: the lack of explicit directional supervision. When the input start and end frames contain weak or ambiguous motion cues—such as a moving object that appears nearly static or centered—the model may struggle to determine a consistent motion direction.

Figure 3 illustrates a limitation of our method when interpolating motion under ambiguous directional cues. In this example, both the start and end frames exhibit only subtle motion signals. Although Bi-AGMI successfully aligns the generated sequence with both endpoints, the inferred motion

trajectory varies across repeated sampling runs under identical conditions.

As shown in Example 1 and Example 2, the upper-layer clouds move in opposite horizontal directions, and the lower-layer clouds also exhibit inconsistent drift. This variation stems from the inherent stochasticity of the diffusion process and the lack of explicit directional supervision, leading to multiple plausible but inconsistent trajectories when motion is underdetermined.

## IV. CONCLUSION

We presented Bi-AGMI, a bidirectional attention-gated motion injection framework for keyframe interpolation, which enhances temporal consistency by directly exchanging attention features across forward and backward denoising paths. Our design enables lightweight and efficient generation while maintaining high perceptual quality.

This work highlights the potential of guided bidirectional diffusion with minimal parameter tuning, offering a scalable solution for high-quality long-interval video interpolation.

## ACKNOWLEDGMENT

The results of this research were obtained from the commissioned research (JPJ012368C05101) by National Institute of Information and Communications Technology (NICT), Japan.

## REFERENCES

[1] H. Feng, Z. Ding, Z. Xia, S. Niklaus, V. Abrevaya, M. J. Black, and X. Zhang, "Explorative inbetweening of time and space," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 378–395, 2024.

[2] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "MCVD: Masked conditional video diffusion for prediction, generation, and interpolation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 23371–23385, 2022.

[3] D. Danier, F. Zhang, and D. Bull, "LDMVFI: Video frame interpolation with latent diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, pp. 1472–1480, 2024.

[4] Z. Huang, Y. Yu, L. Yang, et al., "Motion-aware latent diffusion models for video frame interpolation," *arXiv preprint arXiv:2404.13534*, 2024.

[5] S. Yang, T. Kwon, and J. C. Ye, "ViBiDSampler: Enhancing Video Interpolation Using Bidirectional Diffusion Sampler," *arXiv preprint arXiv:2410.05651*, 2024, under review in ICLR 2025.

[6] X. Wang, B. Zhou, B. Curless, et al., "Generative Inbetweening: Adapting Image-to-Video Models for Keyframe Interpolation," *arXiv preprint arXiv:2408.15239*, 2024.

[7] T. Zhu, D. Ren, Q. Wang, et al., "Generative Inbetweening through Frame-wise Conditions-Driven Video Generation," *arXiv preprint arXiv:2412.11755*, 2024.