AQUA: Aligned Query Fusion for Reference-Unbiased and Temporally Consistent Video Motion Transfer

Jiwoo Park[*1,2] **Sunyoung Jung**[*1] **Yeonkyung Lee**[1] **Tae Eun Choi**[1] **Seong Jae Hwang**[1]

Figure 1. **Teaser.** We present AQUA, an optimization-free motion transfer framework achieving faithful semantic alignment and temporal consistency by mitigating reference bias and flickering.

Abstract

Diffusion models have enabled motion transfer that reflects motion from a reference video while aligning with a given target prompt. While prior methods often require costly model training or fine-tuning, training-free alternatives utilizing self-attention query features have emerged as a flexible solution. However, direct use of reference query features often generates videos with unwanted visual details from the reference and temporal inconsistency. In this paper, we propose AQUA, an optimization-free framework that modulates query features to address these challenges. Specifically, AQUA adaptively balances reference and target query features to remain faithful to the target prompt. Furthermore, AQUA employs a multi-frame guidance to ensure temporal consistency across the generated video.

[*]Equal contribution [1]Department of Artificial Intelligence, University of Yonsei, Seoul, South Korea [2]LG Electronics, Seoul, South Korea. Correspondence to: Seong Jae Hwang <seongjae@yonsei.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Controlling fine-grained motion dynamics is a critical challenge in text-to-video generation. Motion transfer addresses this by generating videos that reflect a target prompt while following the motion of a reference video (Xiao et al., 2024; Jeong et al., 2023). While early methods required computationally expensive fine-tuning or optimization, recent training-free approaches have explored flexible and efficient designs. Notably, Motion by Queries (Atzmon et al., 2024) introduces a simple yet effective strategy that utilizes self-attention queries from a reference video.

Based on this, we analyze the query features in video diffusion models and find that self-attention queries inherently encode crucial spatial details and motion patterns. This observation reveals that while queries are sufficient for guiding motion, their naive usage, termed "direct query injection" introduces two major challenges: reference bias and temporal inconsistency. *Reference bias* is the unintended transfer of reference appearance caused by the spatial information inherently encoded in query features, which leads to deviation from the target prompt. Simultaneously, temporal inconsistency manifests as artifacts stemming from query variations during the stochastic diffusion process, as shown in Fig. 1 (middle row).

To address these, we propose AQUA, an optimization-free framework that utilizes coordinated query features to capture motion-specific information while suppressing irrelevant spatial content. We adaptively fuse reference and target queries using a discrepancy function that quantifies their statistical distribution alignment. Through this distribution-aware fusion, we modulate the influence of each query, effectively mitigating reference bias while preserving essential motion cues. Simultaneously, to ensure temporal consistency, we design a multi-frame guidance mechanism that stabilizes query features during the injection process. By interpolating attention outputs for each frame, our ap-

proach propagates cross-frame information to the queries, effectively mitigating frame-to-frame variance and yielding temporally coherent videos.

2. Related Work

Motion transfer generates a video that follows the motion patterns of a reference while aligning with a target prompt. The core challenge lies in decoupling motion information from the reference appearance to prevent unintended visual influence. While training-based methods such as MotionDirector (Zhao et al., 2024) and DeT (Shi et al., 2025) offer precise control, they are computationally expensive and require substantial training time. In contrast, training-free approaches like MotionClone (Ling et al., 2024) and MotionInversion (Wang et al., 2024) have emerged as efficient alternatives. Specifically, Motion by Queries (Atzmon et al., 2024) utilizes self-attention queries as guidance to enable optimization-free transfer. However, this strategy remains limited by spatial information leakage and query variations, necessitating the modulation of features and a consistent guidance mechanism to ensure high-quality results.

3. Methods

3.1. Analysis of Query in Self-attention

Self-attention query features have been shown to encode rich spatial information (Cao et al., 2023; Alaluf et al., 2024). Motivated by this, we investigate whether the spatial structure in query features persists over time and manifests as coherent motion. To validate this, we visualize the query features using principal component analysis (PCA).

Fig. 2 shows that distinct colors correspond to specific regions in each frame, differentiating objects such as the dog from the surrounding elements. This indicates that the query features capture *spatially* distinct patterns. Furthermore, we observe that the color assignments remain consistent across frames, implying *temporal* coherence in the query features. This observation is illustrated by the t-x slices in Fig. 2, where time progresses along the horizontal axis and the visualized features form smooth and continuous patterns, similar to a reference video. This suggests that the model can capture the temporal dynamics through query features.

We also compare PCA projections of query features with video features to assess their similarity. Table 1 shows that query features exhibit similarity to video features, comparable to that between clean and Gaussian-noised (10%) video features. Although query features show lower similarity under cosine similarity, which is relatively insensitive to noise, other metrics demonstrate stronger alignment. Specifically, both the Mahalanobis distance, which considers feature covariance, and motion fidelity, which captures temporal

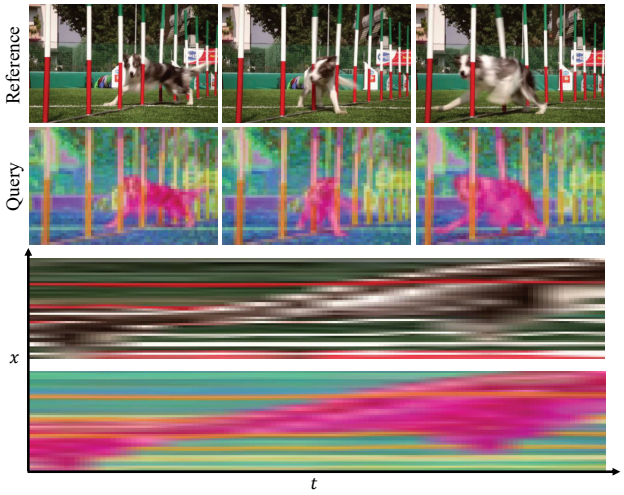


Figure 2. **Query analysis.** Visualization of the reference video, PCA projection query features, and their t-x slices. The query feature maintains consistent spatial representations across time, aligned with the reference video.

Table 1. **Similarity between query and video.** Alignment between query and video features is assessed via cosine similarity, Mahalanobis distance, and motion fidelity (MF).

Compared Entities	Cosine sim ↑	Mahalanobis ↓	MF ↑
Noisy video vs. Video	0.977	0.095	0.831
Query vs. Video	0.853	0.061	0.946

similarity, indicate high similarity between features. These results indicate that query features encode significant structured *spatio-temporal* information from the reference video, making them suitable for motion guidance.

However, naively using query features without considering their properties has critical limitations in motion transfer. First, the spatial appearance from the reference queries dominates, disregarding the target text prompt. This often leads to the unintended transfer of objects or background from the reference video. Second, injecting query features often causes frame-to-frame incoherence due to the gap between the extraction and injection diffusion steps. This leads to flickering artifacts across frames, reflecting degraded temporal coherence in the generated video. Therefore, we introduce a query-based approach that mitigates reference bias and enhances temporal consistency.

3.2. Overall Pipeline: AQUA

We provide an overview of our pipeline, AQUA, as shown in Fig. 3a. To extract query features encoding *spatio-temporal* information, we first apply DDPM inversion (Huberman-Spiegelglas et al., 2024) to the reference video, storing self-attention queries at each diffusion timestep. In the subsequent generation stage, we introduce two key components to address the inherent challenges of motion trans-

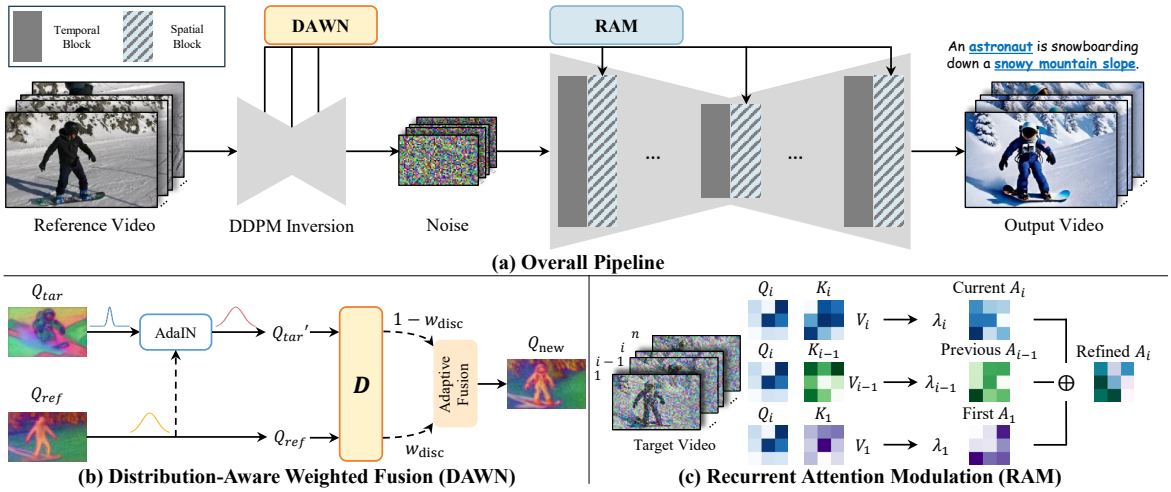


Figure 3. **AQUA Overview.** (a) Query features extracted via DDPM inversion guide video generation. (b) DAWN adaptively fuses reference and target queries via the discrepancy function $\mathcal{D}$. (c) RAM ensures temporal consistency by aggregating attention outputs from the current (i), previous ($i-1$), and first frames.

fer. First, Distribution-Aware Weighted Fusion (DAWN) is performed to inject queries while resolving reference bias (Sec. 3.3). Then, we incorporate Recurrent Attention Modulation (RAM) to mitigate the temporal inconsistency typically introduced by the injection process (Sec. 3.4).

3.3. Distribution-Aware Weighted Fusion

Direct injection often results in reference bias, where irrelevant spatial details from the reference override the target-specific characteristics. To address this issue, we propose Distribution-Aware Weighted Fusion (DAWN), which adaptively fuses reference Q_{ref} and target queries Q_{tar} based on their statistical alignment, as presented in Fig. 3b.

As a preliminary step, we apply Adaptive Instance Normalization (AdaIN) (Huang & Belongie, 2017) to the target query Q_{tar}, producing $Q_{\text{tar}'}$. AdaIN is primarily used to adjust feature distribution in domain adaptation, but we employ AdaIN to harmonize the query features for coherent fusion, which is defined as follows:

$$\begin{aligned} \text{AdaIN}(x, y) &= \sigma(y) \cdot \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y), \\ Q_{\text{tar}'} &= \text{AdaIN}(Q_{\text{tar}}, Q_{\text{ref}}), \end{aligned} \tag{1}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the frame-wise mean and standard deviation. Subsequently, to control the weight of reference and target queries, we introduce a distribution-aware fusion mechanism that utilizes a discrepancy function $\mathcal{D}$ to quantify the statistical distance. The distribution discrepancy function $\mathcal{D}$ is defined as:

$$\begin{aligned} \mathcal{D}(Q_1, Q_2) &= \|\mathbb{E}_{q\sim Q_1}[q] - \mathbb{E}_{q'\sim Q_2}[q']\|_1 \\ &+ \left\| \sqrt{\text{Var}_{q\sim Q_1}[q]} - \sqrt{\text{Var}_{q'\sim Q_2}[q']} \right\|_1 . \end{aligned} \tag{2}$$

Inspired by the 1-Wasserstein distance, this formulation compares the frame-wise mean ($\mathbb{E}_{q\sim Q}[q]$) and standard deviation ($\sqrt{\text{Var}_{q\sim Q}[q]}$).

Utilizing the discrepancy function $\mathcal{D}$, we present an adaptive mechanism that adjusts the weight factor w_{disc} for fusing reference and target queries. When reference and target queries are highly similar, greater weight is assigned to the reference query, as it contains motion information to be transferred. Conversely, when a high divergence is quantified, the target query is emphasized to suppress irrelevant visual content from the reference query. A detailed analysis of the discrepancy function is provided in Supp. Sec. B.2. Based on this design, the adaptive weighting strategy for the fused query Q_{new} is formulated as follows:

$$\begin{aligned} w_{\text{disc}} &= \sigma(\alpha_{\text{disc}} \cdot \mathcal{D}(Q_{\text{tar}'}, Q_{\text{ref}})), \\ Q_{\text{new}} &= (1 - w_{\text{disc}}) \cdot Q_{\text{ref}} + w_{\text{disc}} \cdot Q_{\text{tar}'}, \end{aligned} \tag{3}$$

where σ denotes the sigmoid function and α_{disc} is a scalar scaling factor that controls the sensitivity of the weighting. This computation is performed at every diffusion timestep, allowing dynamic fusion of the reference and target queries.

3.4. Recurrent Attention Modulation

Despite addressing reference bias, the query injection remains susceptible to temporal inconsistency. The gap between the extraction and injection diffusion processes exacerbates variation in the query features during injection. Consequently, injecting such highly unstable query features degrades the temporal coherence of the generated video.

Therefore, we propose Recurrent Attention Modulation (RAM) to improve temporal consistency while retaining motion dynamics, as presented in Fig. 3c. RAM incorporates

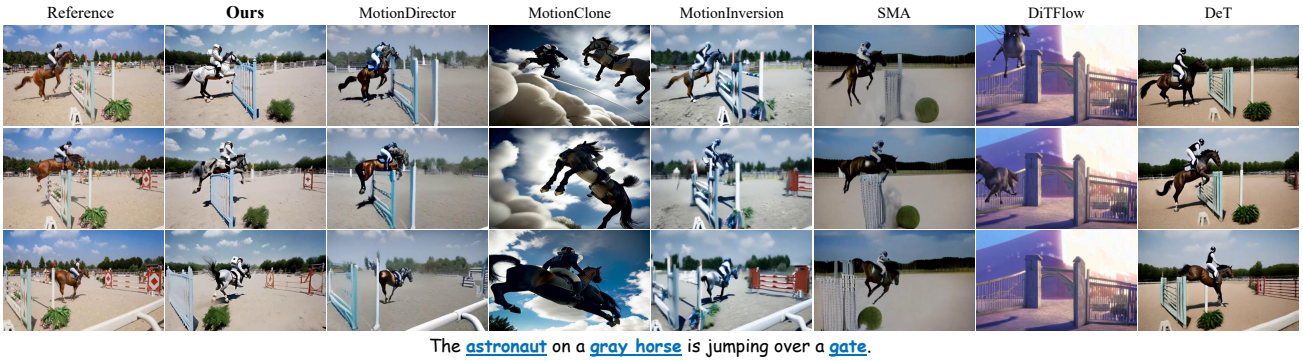


Figure 4. **Qualitative Results.** We provide qualitative results for our method and other baseline models.

information from both the previous and the first frames to refine the attention output for the current frame, $\widetilde{A}_i$, which is computed by $\text{Attn}(Q, K, V) = \text{softmax}\left(QK^\top/\sqrt{d_k}\right)V$. Previous studies (Tan et al., 2025; Lu et al., 2024; Wu et al., 2023) indicate that the first frame $\widetilde{A}_1$ offers global spatial guidance, while the previous frame $\widetilde{A}_{i-1}$ provides short-term temporal continuity. Motivated by these insights, we incorporate temporal context into the attention mechanism and define the multi-frame attention process that produces the final output A_i as follows:

$$\begin{aligned}\widetilde{A}_1 &= \text{Attn}(Q_i, K_1, V_1),\\ \widetilde{A}_{i-1} &= \text{Attn}(Q_i, K_{i-1}, V_{i-1}),\\ \widetilde{A}_i &= \text{Attn}(Q_i, K_i, V_i),\\ A_i &= \lambda_1\widetilde{A}_1 + \lambda_{i-1}\widetilde{A}_{i-1} + \lambda_i\widetilde{A}_i, \end{aligned} \tag{4}$$

where λ_j are weighting coefficients satisfying $\sum_j \lambda_j = 1$ for $j \in \{1, i-1, i\}$, controlling the influence of each frame.

We compute attention outputs independently for each frame and aggregate them using a temporal linear interpolation. This design is motivated by the observation that directly concatenating key-value pairs across frames (i.e., $\text{Attn}(Q_i, [K_i, K_1, K_{i-1}], [V_i, V_1, V_{i-1}])$) (Tewel et al., 2024; Wu et al., 2023) often results in overly static generations and compromises motion fidelity (Fan et al., 2024). In contrast, our method enables the model to leverage spatial information from other frames while preserving motion fidelity. Additional experiments related to motion fidelity can be found in Supp. Sec. D.3. By guiding attention with multi-frame context, our approach mitigates temporal inconsistencies without sacrificing the video generation quality.

4. Experiments

4.1. Experiment Settings

Datasets and Metrics. We use MTBench_HQ, the motion transfer benchmark proposed in DeT (Shi et al., 2025). The dataset consists of 90 reference videos, each paired with 5 different target prompts. We adopt the following metrics for evaluation: CLIP score (EF), DINO-based temporal consistency (TC), and Hybrid Motion Fidelity (HMF), which integrates Motion Fidelity (MF) and Fréchet Distance (Fréchet). For the detailed comparison, we report all three metrics: MF, Fréchet, and HMF. For human evaluation, we conduct a user study involving 30 participants, evaluating three key criteria: editing accuracy, temporal consistency, and motion accuracy. Detailed descriptions of the evaluation and user study are provided in Supp. Sec. C.

Baselines. We compare AQUA with state-of-the-art motion transfer methods, including U-Net-based methods such as MotionDirector (Zhao et al., 2024), SMA (Park et al., 2025), MotionClone (Ling et al., 2024) and MotionInversion (Wang et al., 2024), as well as Diffusion Transformer (DiT)-based methods like DiTFlow (Pondaven et al., 2025) and DeT.

4.2. Qualitative Results

As illustrated in Fig. 4, most baseline models struggle to generate objects corresponding to the target prompt or exhibit severe frame-to-frame variations. In contrast, AQUA generates videos that are both semantically aligned with the prompt and temporally coherent. Notably, our approach remains robust even under highly complex motions (e.g., horse jumping). More qualitative results are provided in the supplementary materials.

4.3. Quantitative Results

Table 2 compares AQUA with existing motion transfer approaches. AQUA outperforms prior works across most quantitative metrics. In particular, it achieves superior scores in EF, MF, Fréchet Distance, and HMF, demonstrating strong motion fidelity and prompt alignment while effectively mitigating reference bias. Furthermore, the high scores in the human evaluation indicate that AQUA's outputs closely align with human preferences. While existing DiT-based methods often exhibit high temporal consistency but limited motion fidelity, AQUA maintains comparable performance without compromising motion. AQUA is also highly efficient, requiring the lowest computational cost without additional

Table 2. **Quantitative comparison.** We conduct our experiments on the MTBench_HQ. The evaluation consists of three aspects: (1) Automatic metrics, referring to algorithmically computed measures that assess prompt alignment, temporal consistency, and motion fidelity; (2) Human evaluation, providing subjective assessment of the same criteria; and (3) Efficiency metrics, including computation time and peak memory usage for generating a motion-transferred video from a reference video.

Model	Venue	Automatic Metrics					Human Evaluation			Efficiency Metrics	
		EF ↑	TC ↑	MF ↑	Fréchet ↑	HMF ↑	EF ↑	TC ↑	MF ↑	Time (s)	Memory (GB)
MotionDirector	ECCV'24	31.6	89.1	62.5	91.6	77.1	62.4	65.8	59.1	175	10.16
SMA	AAAI'25	30.7	83.8	71.5	92.1	81.8	71.3	60.9	66.7	1125	25.34
MotionInversion	Arxiv'24	30.0	85.6	89.0	91.4	90.2	59.8	63.6	75.2	476	9.16
MotionClone	ICLR'25	31.3	84.3	78.7	90.0	84.4	67.5	51.4	43.8	82	13.38
DiTFlow	CVPR'25	31.4	**94.6**	59.5	91.8	75.6	80.2	81.6	42.1	87	11.27
DeT	ICCV'25	29.8	91.8	56.9	92.0	74.5	56.0	84.0	91.1	2014	25.28
AQUA		**32.7**	90.1	**91.0**	**92.5**	**91.7**	**92.6**	**87.0**	**93.6**	**48**	**7.25**

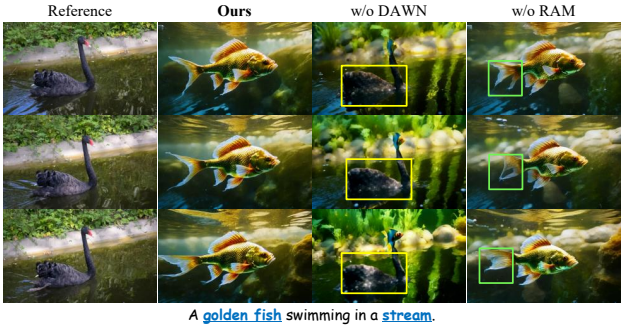


Figure 5. **Ablation study.** The yellow box indicates reference bias, while the green box highlights temporal inconsistency.

Table 3. **Ablation results.**

Exp.	DAWN	RAM	EF	TC	MF	Fréchet	HMF
1 (MbQ)			30.2	85.7	86.9	91.4	89.2
2		✓	30.9	89.0	89.4	91.9	90.6
3	✓		32.3	87.8	88.8	92.2	90.5
4	✓	✓	**32.7**	**90.1**	**91.0**	**92.5**	**91.7**

training or optimization. These results show the comprehensive superiority and practical effectiveness of our approach.

4.4. Ablation Study

We conduct ablation studies by comparing AQUA with a direct query injection method, Motion by Queries (MbQ) (Atzmon et al., 2024), to evaluate the contribution of each component. Table 3 shows that DAWN improves the CLIP score, while RAM specifically enhances temporal consistency. As shown in Fig. 5, the removal of DAWN leads to the leakage of the reference appearance, where the shape of the swan from the reference video appears in the generated frames. Conversely, RAM reduces object inconsistencies, mitigating variations in the size and shape of the fish that occur in its absence. These experiments validate that AQUA effectively addresses reference bias and temporal inconsistency, further enhancing motion fidelity in the generated video.

Table 4. **Quantitative comparison of distributional metrics.** We compare alternative distributional metrics within DAWN.

Method	EF	TC	MF	Fréchet	HMF
KL-Divergence (Pérez-Cruz, 2008)	32.1	88.7	88.7	92.1	90.4
MMD (Li et al., 2015)	32.1	88.6	88.6	92.1	90.4
AQUA	**32.7**	**90.1**	**91.0**	**92.5**	**91.7**

4.5. Analysis of Distributional Metrics

We justify our discrepancy function $\mathcal{D}$ by comparing it with the KL-divergence (Pérez-Cruz, 2008) and Maximum Mean Discrepancy (MMD) (Li et al., 2015), which are widely adopted metrics for measuring distances between probability distributions. As shown in Table 4, AQUA consistently outperforms other distribution functions across overall metrics. This indicates that our proposed $\mathcal{D}$ captures the spatial divergence in query features more robustly than standard density or kernel-based measures, effectively balancing reference motion and target appearance.

5. Conclusion

We propose AQUA, an optimization-free motion transfer framework ensuring faithful prompt alignment and temporal consistency. Our work identifies limitations in direct query injection, such as reference bias and frame-level inconsistency, despite the effectiveness of query features in capturing spatio-temporal cues. By leveraging adaptive query fusion and an inter-frame attention mechanism, AQUA enables semantically aligned and temporally coherent motion transfer. This approach establishes a highly efficient, training-free motion transfer paradigm.

Limitations. While AQUA effectively improves semantic alignment and temporal consistency, query-level modulation primarily operates on global spatio-temporal cues. As a result, fine-grained local dynamics, such as subtle motion, may remain challenging, as shown in Fig. 17. Extending AQUA with fine-grained local motion guidance is therefore a promising direction for further improving motion transfer.

References

Alaluf, Y., Garibi, D., Patashnik, O., Averbuch-Elor, H., and Cohen-Or, D. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Atzmon, Y., Gal, R., Tewel, Y., Kasten, Y., and Chechik, G. Motion by queries: Identity-motion trade-offs in text-to-video generation. *arXiv preprint arXiv:2412.07750v3*, 2024.

Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., and 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.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., and Shan, Y. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

Chung, J., Hyun, S., and Heo, J.-P. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8795–8805, 2024.

Fan, J., Xue, H., Zhang, Q., and Chen, Y. Refdrop: Controllable consistency in image or video generation via reference feature guidance. *Advances in Neural Information Processing Systems*, 37:33602–33637, 2024.

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Hertz, A., Voynov, A., Fruchter, S., and Cohen-Or, D. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4775–4785, 2024.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Huberman-Spiegelglas, I., Kulikov, V., and Michaeli, T. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12469–12478, 2024.

Jeong, H., Park, G. Y., and Ye, J. C. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. *arXiv preprint arXiv:2312.00845*, 2023.

Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., and Rupprecht, C. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015.

Ling, P., Bu, J., Zhang, P., Dong, X., Zang, Y., Wu, T., Chen, H., Wang, J., and Jin, Y. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024.

Lu, Y., Liang, Y., Zhu, L., and Yang, Y. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024.

Park, G. Y., Jeong, H., Lee, S. W., and Ye, J. C. Spectral motion alignment for video motion transfer using diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6398–6405, 2025.

Pérez-Cruz, F. Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pp. 1666–1670. IEEE, 2008.

Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., and Pizzati, F. Video motion transfer with diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22911–22921, 2025.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shi, Q., Wu, J., Bai, J., Zhang, J., Qi, L., Li, X., and Tong, Y. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. *arXiv preprint arXiv:2503.17350*, 2025.

Si, C., Huang, Z., Jiang, Y., and Liu, Z. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4733–4743, 2024.

Tan, J., Yu, H., Huang, J., Xiao, J., and Zhao, F. Freepca: Integrating consistency information across long-short frames in training-free long video generation via principal component analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27979–27988, 2025.

Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., and Atzmon, Y. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43 (4):1–18, 2024.

Wang, L., Mai, Z., Shen, G., Liang, Y., Tao, X., Wan, P., Zhang, D., Li, Y., and Chen, Y. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*, 2024.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

Xiao, Z., Zhou, Y., Yang, S., and Pan, X. Video diffusion models are training-free motion interpreter and controller. *arXiv preprint arXiv:2405.14864*, 2024.

Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., and Huang, T. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Yatim, D., Fridman, R., Bar-Tal, O., Kasten, Y., and Dekel, T. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8466–8476, 2024.

Zhang, L., Rao, A., and 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.

Zhao, R., Gu, Y., Wu, J. Z., Zhang, D. J., Liu, J.-W., Wu, W., Keppo, J., and Shou, M. Z. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024.

Figure 6. **Qualitative Results.** We present AQUA, a training-free motion transfer method for text-to-video diffusion models that enhances semantic reflection and temporal consistency. Direct query injection introduces unintended reference content, such as a mountain background, a backpack artifact (left), and a car (right), along with abrupt visual artifacts like a green object (left) and a purple region (right). In contrast, AQUA generates temporally coherent videos, faithfully aligned with the target prompt.

A. Code and Website

The code for AQUA is included in the supplementary materials as `code.zip`. To facilitate accessibility, we present our results on the project page, which also includes a brief presentation video. The project page is available at: https://aqua-aaai26.github.io/aqua/.

B. Method Details

B.1. Detailed Procedure for Query Analysis

We investigate the self-attention query features in the context of motion transfer. We first extract query features through the diffusion inversion process. Specifically, we select the 25th timestep out of 50 total steps to obtain representative query features. To evaluate and visualize the extracted features, we apply Principal Component Analysis (PCA) to reduce their dimensionality to three dimensions as follows:

$$\mathrm{Q} \in \mathbb{R}^{F \times HW \times D} \xrightarrow{\text{PCA}} \mathbb{R}^{F \times HW \times 3}, \tag{1}$$

where Q denotes the self-attention query, F is the number of frames, H, W are the height and width, and D is the feature dimension of the query. This dimensionality reduction is applied along the query dimension D, enabling visualization of the features in a three-dimensional space similar to RGB space. The resulting visualizations reveal that the query features exhibit structured patterns resembling the spatial layout of the reference video.

B.2. Details of Discrepancy function $\mathcal{D}$

We analyze the distributional discrepancy between reference and target query features. For this analysis, the target query is extracted during video generation using the target prompt, without injecting the reference query. Fig. 7 presents a comparison of the temporally averaged query feature distribution for the Reference, Target 1, and Target 2. Each distribution is represented as a probability density function (PDF) and overlaid within a single plot for direct visual comparison. This visualization facilitates an intuitive assessment of the degree of alignment and divergence among the three query distributions. Target 2, prompted with 'penguin', shows a larger distributional discrepancy from the Reference compared to Target 1, which is prompted with 'deer'. This is attributed to the substantial difference in object appearance and background layout—while the Reference and Target 1 both depict forest scenes, Target 2 features a non-forest environment. These observations suggest that the discrepancy score captures spatial divergence between reference and target query features.

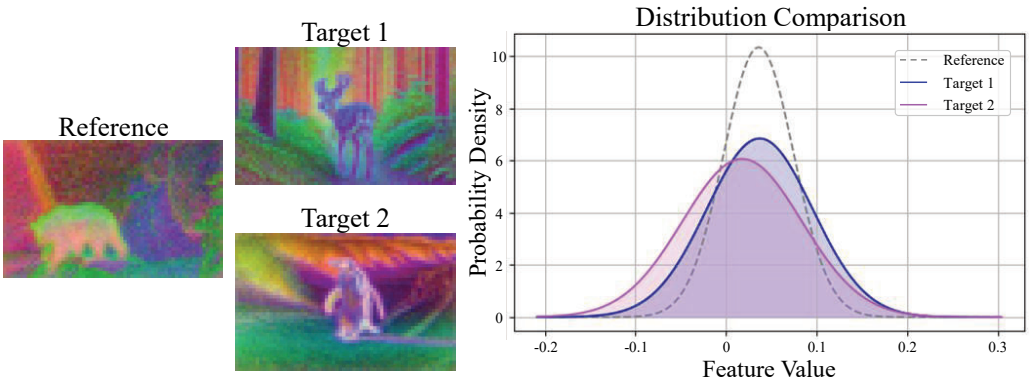


Figure 7. **Analysis of query feature distribution.** Probability density functions (PDFs) of a reference query and two different target queries, illustrating their distribution.

Figure 8. **Analysis of α_{disc} in DAWN.** Qualitative results under different α_{disc} settings in DAWN. The comparison includes settings of 0.1 (default in our method), 0.5, and 0.8.

B.3. Analysis of Hyperparameters

B.3.1. Hyperparameter α_{DISC}

We investigate how varying the scaling scalar factor α_{disc} in DAWN affects the results. Other experimental settings are kept fixed, with α_{disc} set to 0.1 by default in all previously reported results. This hyperparameter serves as a sensitivity control that determines how strongly the adaptive weighting mechanism responds to distribution discrepancies. A higher value of α_{disc} increases the influence of these discrepancies, resulting in more dynamic weight adjustments, whereas a lower value leads to more stable behavior. As illustrated in Fig. 8, when α_{disc} increases, the stronger influence of the discrepancy-based

Figure 9. **Analysis of λ in RAM.** Qualitative results under different λ configurations. (a) Setting $\lambda_i = 0.5$ causes artifacts, (b) $\lambda_i = 0.8$ yields stable results and visually coherent outputs. Here, i denotes the current frame.

weights hinders faithful motion reproduction from the reference video. This suggests that overly dynamic weight modulation can impair the incorporation of reference video information, resulting in insufficient motion guidance.

B.3.2. HYPERPARAMETER λ_i

Furthermore, we empirically analyze the effect of the weighing factors λ_1, λ_{i-1}, and λ_i in RAM. These hyperparameters control the contributions of the first frame, previous frame, and current frame, respectively. Our results show that assigning a high weight to the current frame λ_i is essential for avoiding visual artifacts in the generated outputs. To balance contextual input from the first and previous frames, we set both λ_1 and λ_{i-1} to 0.1 in our experiments. Fig. 9 presents qualitative results with a reduced λ_i value of 0.5. The results suggest that excessive reliance on contextual information from earlier frames degrades visual quality.

C. Experimental Details

C.1. Dataset

Previous studies on motion transfer have lacked a standardized benchmark dataset. Most existing works generate videos from general video datasets (Xu et al., 2018; Pont-Tuset et al., 2017) using independently selected and unconstrained prompts, which hinders objective evaluation. To address this limitation, DeT (Shi et al., 2025) introduces and publicly releases a benchmark dataset specifically for motion transfer tasks. Therefore, we adopt this dataset to evaluate all baseline models under consistent conditions, enabling fair and objective comparisons. The dataset consists of 90 reference videos, each paired with 5 different target prompts.

C.2. Baselines

We compare our method with state-of-the-art motion transfer methods, including U-Net-based methods such as MotionDirector (Zhao et al., 2024), SMA (Park et al., 2025), MotionClone (Ling et al., 2024) and MotionInversion (Wang et al., 2024), as well as Diffusion Transformer (DiT)-based methods like DiTFlow (Pondaven et al., 2025) and DeT.

C.3. Metrics

We follow the evaluation protocol of DeT (Shi et al., 2025) and report three categories of metrics: (1) CLIP score (EF) to measure text-video alignment; (2) DINO-based temporal consistency (TC) to assess frame-wise coherence; and (3) Hybrid Motion Fidelity (HMF) to quantify motion transfer quality. Since HMF integrates Motion Fidelity (MF) and Fréchet Distance (Fréchet), we additionally report MF and Fréchet for detailed analysis. For human evaluation, we conduct a user study with 30 participants, assessing editing accuracy, temporal consistency, and motion accuracy.

C.3.1. CLIP SCORE (EF)

For the CLIP score, we use the CLIP-ViT-B/32 model (Radford et al., 2021). This metric evaluates the semantic alignment between the generated video and the target text prompt. Since a video consists of multiple frames, we compute the similarity between each frame and the text individually and then average the results across all frames, as follows:

$$\text{CLIP Score} = \frac{1}{N}\sum_{i=1}^{N}\cos(\phi_{\text{frame}}(f_i), \phi_{\text{text}}(T))], \tag{2}$$

where $\phi_{\text{frame}}(\cdot)$ and $\phi_{\text{text}}(\cdot)$ denote the CLIP encoders for frame and text, respectively, and $\cos(\cdot,\cdot)$ indicates cosine similarity.

C.3.2. TEMPORAL CONSISTENCY (TF)

We evaluate temporal consistency using DINO (Caron et al., 2021)-based feature correspondence, which measures the similarity of visual representations across consecutive frames. Specifically, we employ the ViT-B/16 variant of the DINO model to extract frame-level representations. Given a video with N frames $\{f_1, f_2, \ldots, f_N\}$, we obtain features for each frame as $\text{DINO}(f_i)$. The temporal consistency (TC) score is then calculated as the average cosine similarity between features of adjacent frames:

$$\text{TC} = \frac{1}{N-1}\sum_{i=1}^{N-1}\cos\left(\text{DINO}(f_i), \text{DINO}(f_{i+1})\right). \tag{3}$$

C.3.3. HYBRID MOTION FIDELITY (HMF)

DeT (Shi et al., 2025) addresses the limitations of conventional motion fidelity metrics by proposing a new metric called Hybrid Motion Fidelity (HMF). The original motion fidelity metric, introduced in SMM (Yatim et al., 2024), evaluates the similarity between tracklets from different videos using a tracking method (Karaev et al., 2024). For each tracklet, the most similar counterpart is identified via Chamfer Distance, and the similarity of their local velocity directions is computed as follows:

$$\bar{c}_n = \frac{1}{T-1}\sum_{t=1}^{T-1}\cos\left(\Delta\mathcal{T}_i^n(t), \Delta\mathcal{T}_j^n(t)\right), \tag{4}$$

where $\Delta\mathcal{T}(t) = \mathcal{T}(t+1) - \mathcal{T}(t)$ denotes the local velocity vector at time t. This formulation quantifies motion fidelity based on the directional similarity between matched tracklets. In addition, DeT incorporates the Fréchet Distance to compare the overall shape and temporal alignment of trajectories, as follows:

$$\text{Fréchet Distance} = \alpha e^{-d_F(\mathcal{T}_i^n, \mathcal{T}_j^n)}. \tag{5}$$

Unlike cosine similarity, which captures only directional consistency, the Fréchet Distance accounts for both spatial alignment and the temporal ordering, providing a finer-grained comparison between motion patterns. The final HMF score is defined as a weighted combination of the Fréchet distance term and the cosine similarity metric, formulated as:

$$\text{HMF}(\mathcal{T}_i, \mathcal{T}_j) = \frac{1}{N}\sum_{n=1}^{N}\left[\alpha e^{-d_F(\mathcal{T}_i^n, \mathcal{T}_j^n)} + (1-\alpha)\bar{c}_n\right], \tag{6}$$

where α is a weighting coefficient that balances the contribution of trajectory shape similarity (via Fréchet distance) and local velocity similarity (via cosine similarity).

C.4. Implementation details

To demonstrate the applicability of our method, we apply it to various video diffusion models. These include U-Net-based models such as VideoCrafter2 (Chen et al., 2024) and AnimateDiff (Guo et al., 2023), as well as the DiT-based model HunyuanVideo (Kong et al., 2024). Unless otherwise specified, all main experiments are conducted using VideoCrafter2.

We set the number of inference timesteps to 50 and inject query features at 40% of the total diffusion steps. To perform the injection, we extract reference query features using the DDPM inversion (Huberman-Spiegelglas et al., 2024), which addresses the error accumulation issues of DDIM inversion. Specifically, the query features are extracted from all self-attention layers within the spatial attention modules, excluding those in the temporal attention.

Figure 10. **Validation of our discrepancy-based weighting.** Comparison of three query injection methods: *Direct*, which directly injects the reference query; *Balance*, which applies a fixed 0.5 weighting between reference and target queries; and *Ours*, which employs adaptive weighting based on distribution discrepancy between the two queries.

For evaluation, each video in the MTBench_HQ dataset is preprocessed to 32 frames, from which 16 frames are used to construct the reference video. All models use this reference video as input and are required to generate the corresponding 16 output frames. Additionally, for consistency, all efficiency metrics are measured on an RTX A6000 GPU across all models.

C.5. User Study

Since there are no universally accepted objective metrics for evaluating motion transfer in videos, prior works (Park et al., 2025; Ling et al., 2024) have predominantly relied on human evaluation. Following this practice, we conduct a user study to support the superiority of our method through various experiments. Our user study format is based on that of SMA (Park et al., 2025), and includes three criteria: edit fidelity, temporal consistency, and motion fidelity.

- **Edit Fidelity** assesses how accurately the video reflects the target text prompt, specifically whether the described subjects, actions, and attributes are visually represented.
- **Temporal Consistency** measures the smoothness and coherence of visual transition over time throughout the video.
- **Motion Fidelity** evaluates the similarity of the motion in the generated video to that in the reference video.

These are evaluated based on the following questions: (1) Edit Fidelity: *How well does the content of the video align with the given text prompt?* (2) Temporal Consistency: *Does the video maintain natural continuity in motion, appearance, and background over time?* (3) Motion Fidelity: *How closely does the overall motion in the generated video resemble the reference motion?* Participants rate each aspect on a 1-to-5 Likert scale, and we aggregate these scores to compare the performance of different baseline models.

Table 5. **Comparison with other query injection methods.** We provide quantitative results of our discrepancy-based adaptive weighting against other query injection methods.

Method	EF	TC	MF	Fréchet	HMF
Direct	30.2	85.7	86.9	91.4	89.2
Balance	31.6	87.8	89.3	92.0	90.6
AQUA	**32.7**	**90.1**	**91.0**	**92.5**	**91.7**

Table 6. **Quantitative evaluation of DAWN variants.** We evaluate DAWN with and without AdaIN and the discrepancy function $\mathcal{D}$.

Exp.	AdaIN	$\mathcal{D}$	EF	TC	MF	Fréchet	HMF
1	✓		31.4	87.8	88.9	92.1	90.5
2		✓	32.6	94.7	67.3	93.0	80.1
3	✓	✓	32.7	90.1	91.0	92.5	91.7

Figure 11. **Qualitative comparison of DAWN variants.** We provide qualitative results for three configurations: (a) our full method using both AdaIN and the discrepancy function $\mathcal{D}$, (b) a variant using AdaIN with fixed-weight balancing instead of $\mathcal{D}$, and (c) a variant using $\mathcal{D}$ without AdaIN.

D. Additional Results

D.1. Comparison with DAWN

We compare our proposed method, DAWN, against two alternative approaches: direct injection and fixed-weight balancing. In the fixed-weight balancing strategy, reference and target queries are combined with a fixed ratio of 0.5 (Chung et al., 2024).

As shown in Fig. 10, direct injection leads to noticeable reference bias, with the bus from the reference video appearing on the road. Fixed-weight balancing partially alleviates this issue by generating a boat on the lake, as specified in the target prompt, but residual elements from the bus and road are still visible. In contrast, DAWN adaptively adjusts the contribution of each query based on a distributional discrepancy function, enabling more precise motion transfer that reflects the target appearance while accurately following the reference motion. Table 5 provides quantitative results of improved prompt alignment and temporal consistency.

D.2. Additional experiments with DAWN

We provide additional experiments on DAWN in Fig. 11. As shown in Fig. 11, we compare (i) fixed-weight balancing with AdaIN and (ii) our method without AdaIN. In the case of fixed-weight balancing, applying AdaIN still results in reference

Table 7. **Quantitative results of RAM and Concatenated attention method.** We present a comparative analysis with an existing attention method to validate the effectiveness of our proposed RAM approach.

Method	EF	TC	MF	Fréchet	HMF
Concatenated	31.2	87.2	86.8	92.1	89.0
AQUA	**32.7**	**90.1**	**91.0**	**92.5**	**91.7**

Figure 12. **Qualitative results of RAM and Concatenated attention method.** We compare videos generated by our method, RAM, and the key-value concatenated attention approach.

bias, generating a car that resembles the reference video instead of the 'retro bus' described in the target prompt. Quantitative results in Table 6 further show that the fixed-weighting approach exhibits minimal performance difference. In contrast, our adaptive fusion mechanism enables accurate motion transfer while effectively suppressing irrelevant appearance information from the reference.

We also investigate the effect of removing AdaIN from our method. Although its exclusion improves semantic alignment with the target prompt compared to fixed-weighting, it causes the generated motion to become overly static. This is reflected in higher scores for temporal consistency (TC) and Fréchet Distance, but notably lower scores in motion fidelity (MF), indicating that motion diversity is lost. These findings suggest that AdaIN plays a crucial role in harmonizing feature distributions between the reference and target queries. Without AdaIN, discrepancy computation becomes unstable, which in turn destabilizes the adaptive weighting function, ultimately leading to static video generation.

D.3. Comparison with RAM

We compare our proposed module, RAM, which enhances temporal consistency in video generation, with a widely used training-free method (Tewel et al., 2024; Hertz et al., 2024) that concatenates key-value pairs from adjacent frames. The comparison is conducted under the same experimental setup as the main experiments, using the MTBench_HQ benchmark.

As shown in Table 7, RAM consistently achieves superior performance in both temporal consistency and motion fidelity metrics. Fig. 12 further illustrates these findings: the key-value concatenation method yields static motion, such as a deer standing still, whereas our method successfully follows the motion in the reference video. These results support the previous findings (Fan et al., 2024) that naive key-value concatenation not only underperforms in temporal consistency but also degrades motion fidelity. In contrast, RAM improves temporal consistency and motion fidelity, while also achieving higher Fréchet Distance scores.

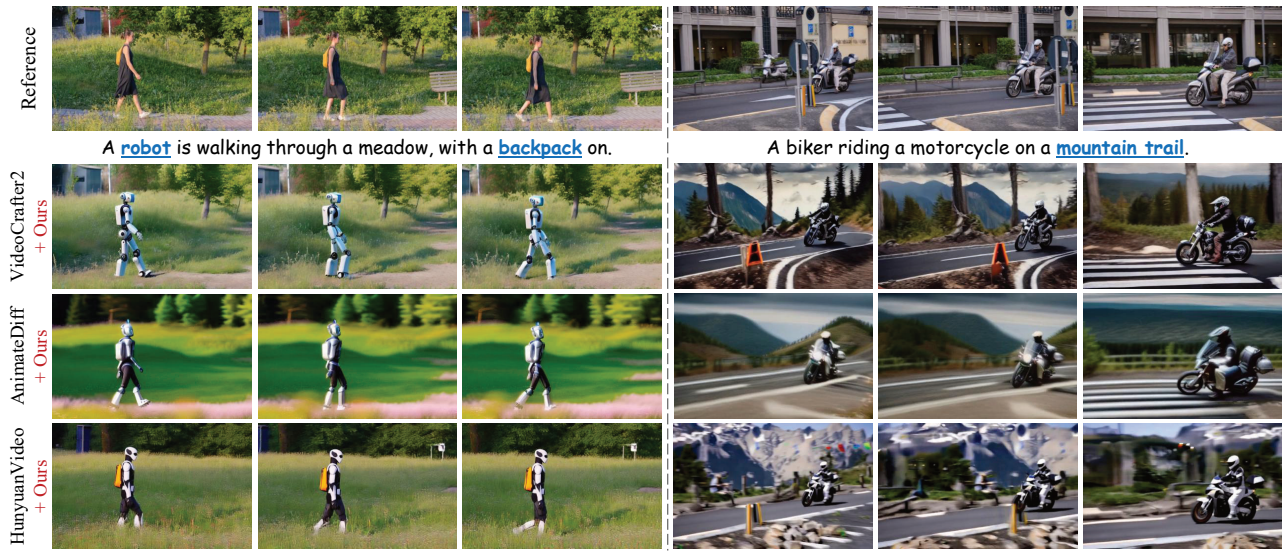


Figure 13. **Application with diverse pretrained models.** We demonstrate the broad applicability of our method by applying it to various pretrained video generation models.

D.4. Generalizability of AQUA

We evaluate AQUA on additional video generative models, including AnimateDiff (Guo et al., 2023) and HunyuanVideo (Kong et al., 2024). As shown in Fig. 13, AQUA enables motion transfer across these architectures without model-specific training, suggesting that our approach generalizes beyond the base model used in previous experiments.

D.5. More Qualitative results

We provide additional qualitative results to demonstrate the applicability of our method across diverse video scenarios. Fig. 14 shows extended comparisons with the baseline models used in the main experiments. Even in videos with rapid motion, our method consistently achieves motion transfer while maintaining alignment with the target prompt. Fig. 15 further presents diverse scenes generated by our method, supporting its robustness across various contexts. In addition, Fig. 16 validates the generalizability of our method by applying it to other pretrained video generation models.

D.6. Failure Case

Like existing motion transfer works, our method faces challenges with subtle movements. As analyzed in Fig. 17, while overall facial dynamics and wrinkles are successfully transferred, precise lip motions are occasionally missed. Our query modulation primarily focuses on overall motion dynamics, which inherently limits its capacity to represent fine-grained local details. This representational bottleneck leads to the observed difficulties in handling subtle movements, aligning with the inherent limitations of current video diffusion models in capturing fine-grained details (Zhang et al., 2023; Si et al., 2024). Consequently, exploring highly precise, fine-grained motion transfer stands as a crucial direction for future research.

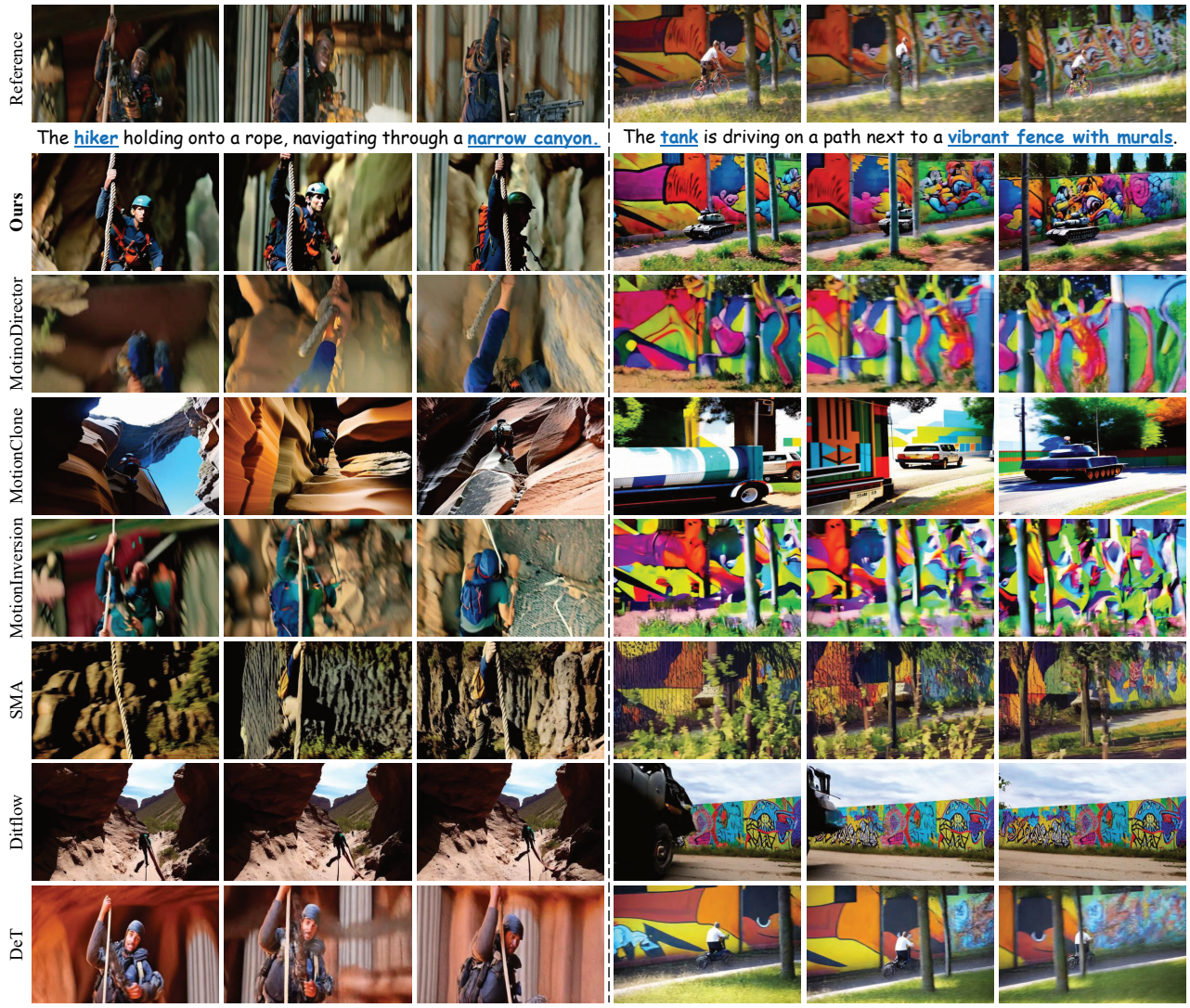


Figure 14. **Additional qualitative comparison results.** We provide more qualitative results for both the baseline models and our method.

[Reference]

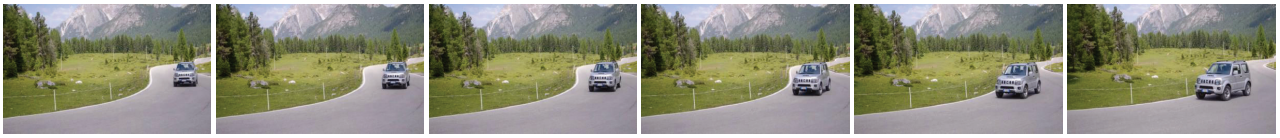
A chicken is walking on the grass near a stream.

A peacock is walking on the grass near a lake.

[Reference]

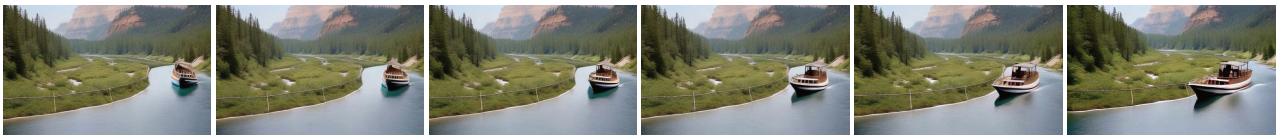
A sleek Formula One car driving on a circuit, navigating a turn.

A black SUV driving on a mountain pass, navigating a turn.

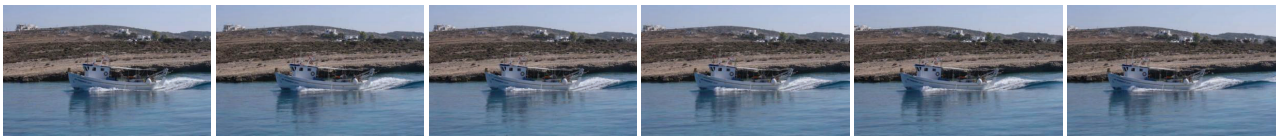


[Reference]

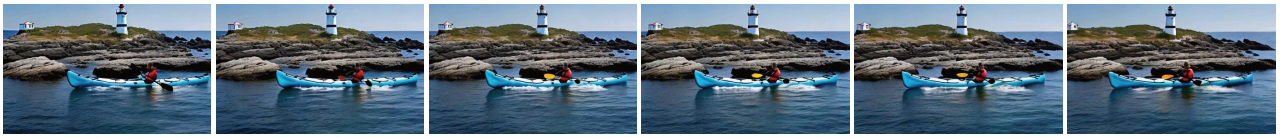
A bus driving on a winding road through a valley filled with wildflowers.



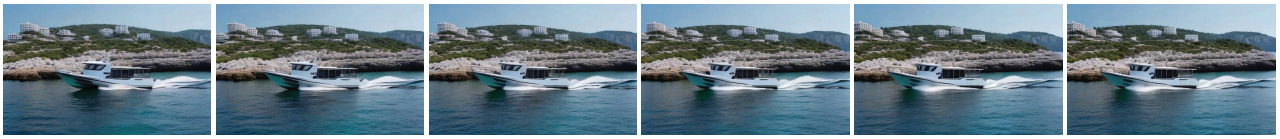
A boat driving on a winding river through a serene canyon.



[Reference]



A kayak traveling on a serene sea with a backdrop of a rocky shore and a few lighthouses.



A motorboat traveling on a quiet bay with a rocky shore lined with modern seaside buildings.

Figure 15. **More qualitative results of our method, AQUA.** Additional qualitative results of our method on the MTBench_HQ dataset.

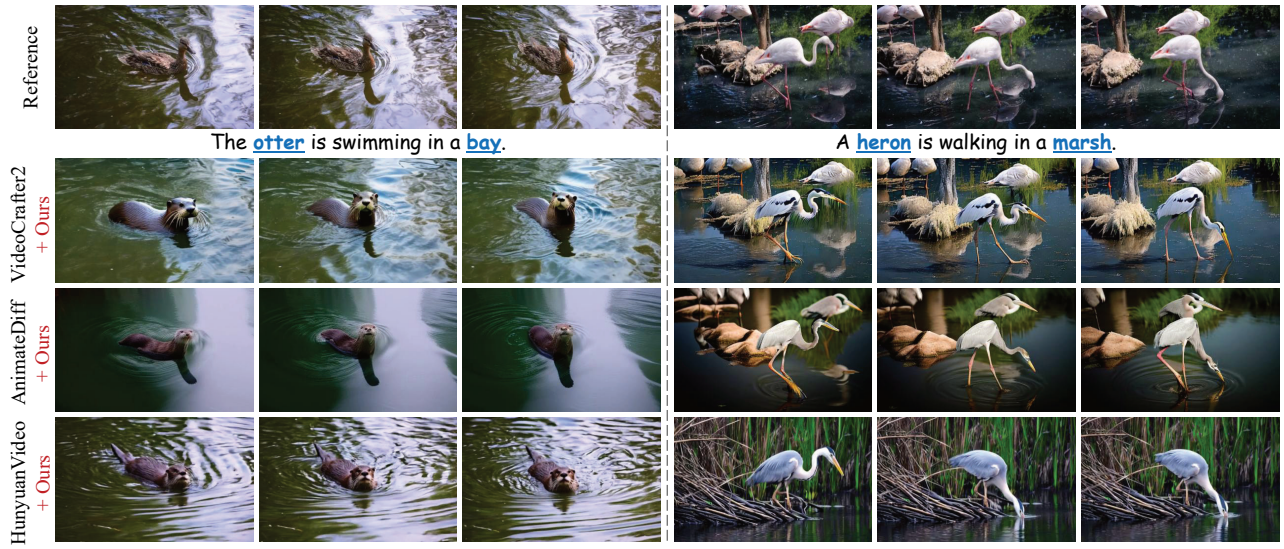


Figure 16. **Additional qualitative results in diverse pretrained models.** We provide more generation results applying to various pretrained video generative models.

Figure 17. **Failure case.** Qualitative results demonstrating a challenging scenario for AQUA : capturing subtle motions.