Identity-Motion Trade-offs in Text-to-Video via Query-Guided Attention Priors

Anonymous ICCV submission

Paper ID *****

Abstract

Text-to-video diffusion models have shown remarkable progress in generating coherent video clips from textual descriptions. However, the interplay between motion, structure, and identity representations in these models remains under-explored. Here, we investigate how self-attention query (Q) features simultaneously govern motion, structure, and identity – revealing these features as key structural priors that control video generation. Our analysis shows that Q affects not only layout, but that during denoising Q also has a strong effect on subject identity, making it hard to transfer motion without the side-effect of transferring identity. Understanding this dual role enabled us to control query feature injection (Q injection) and demonstrate two applications: (1) a zero-shot motion transfer method – implemented with VideoCrafter2 and WAN 2.1 – that is 10× more efficient than existing approaches, and (2) a training-free technique for consistent multi-shot video generation, where characters maintain identity across multiple video shots while Q injection enhances motion fidelity.

1. Introduction

Video generation from text is at the forefront of generative AI. Despite progress in controlling entities in video, major challenges remain: generating natural, engaging motion and preserving consistent identity throughout the video. These goals often form a trade-off – preserving consistency is easier with limited motion, while increased motion makes consistency harder as entity appearance changes. A key challenge is understanding how motion and identity are represented in video models and how to control them effectively.

This limited understanding hinders downstream applications. While many motion transfer approaches [17, 28, 30, 32] rely on tuning or test-time optimization, there is growing interest in inference-time methods that exploit internal representations rather than computational scale, similar to text-to-image layout transfer through feature manipulation [1, 3]. Better model understanding could lead to further progress in this direction. As another example, consider consistent characters in multi-shot video generation, where the goal is to preserve consistency of character identity and appearance across shots. Image-based models tackle this through feature-sharing, but applying the same ideas to video leads to loss of motion because the shared features encode both identity *and* motion.

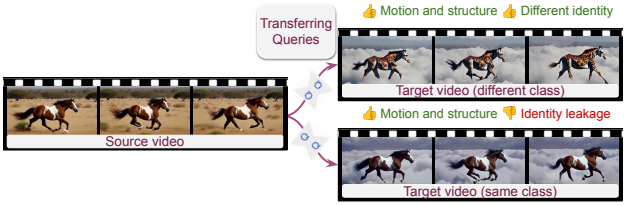

Figure 1. Our analysis reveals differences in Q-injection between text-to-video and text-to-image models. One key observation is that in text-to-video models, Q injection transfers both motion and structure, but suffers from identity leakage when source and target share the same subject.

To understand motion and identity representation in video models, we draw analogies from text-to-image (T2I) models which are better understood. Motion can be viewed as "3D-shape" in the tensor defined by frame sequences, making it natural to examine shape/structure representation in image models. Previous works showed diffusion-based T2I models establish layout in early steps [21]. Studies [1, 3, 26] found self-attention queries (Q) encode structural layout priors and injecting queries between images during generation "copies" shape while preserving appearance. As motion is the video equivalent of structure, we investigate the relationship between query vectors, motion, and identity in video generation.

We conduct an empirical analysis revealing that unlike image models, Q in video-generation models affects both motion and identity, and videos require more denoising steps than images to capture motion patterns. We leverage this insight for two applications: motion transfer and consistent multi-shot video generation.

For motion transfer, injecting Q features from a source video during denoising enables zero-shot motion transfer without fine-tuning. Our pipeline achieves quality close to leading methods while being 10× more efficient than exist-

ing approaches, demonstrating that understanding internal representations can outperform scaling computation.

For consistent multi-shot video generation, we build on insights from multi-shot image generation [26] using extended attention shared between video shots. While Q injection from unconstrained generation preserves motion diversity, video generation requires more Q injection steps than images, causing identity leakage that compromises shot-to-shot consistency. We address this with two phases: (1) Q-Preservation – maintains motion structure using unconstrained Q values (2) Q-Flow – preserves feature flow maps to avoid identity leakage in later steps.

Our main contributions: 1) A systematic analysis of Q-features as a structural prior in text-to-video diffusion models, revealing their dual role in encoding both motion and identity, with effects persisting longer into the denoising process. 2) "*Motion by Queries*", an efficient zero-shot motion transfer approach for both UNeT and Diffusion-Transformer architectures. 3) A training-free method for consistent multi-shot video generation balancing character consistency and motion quality.

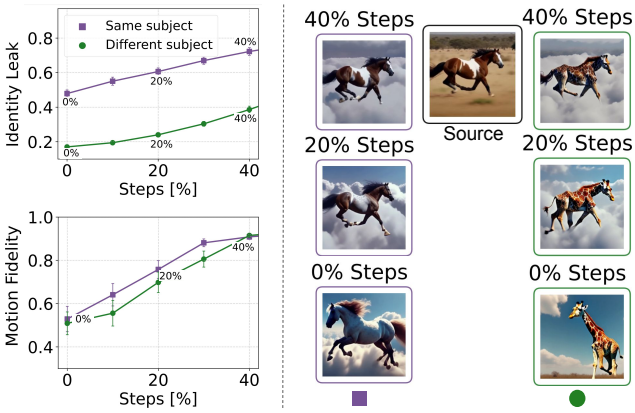

Figure 2. Same-class transfer shows identity leakage increasing with Q injection (purple), while cross-class transfer (green) maintains identity separation at 40% injection where motion quality is preserved. Data from [28].

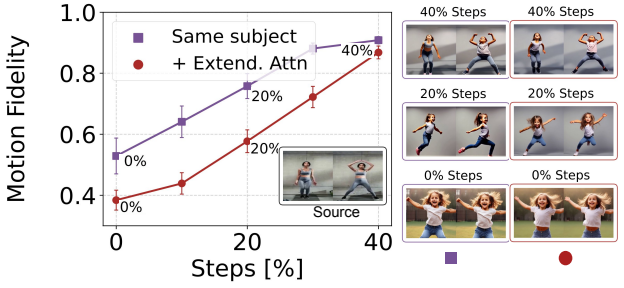

Figure 3. Extended attention reduces motion, requiring longer Q-injection to recover it – consequently increasing identity leakage.

2. Analysis: Query features in text-to-video generation

To study the role of Q-features in text-to-video (T2V) models, we design an injection experiment. The idea is simple: we take a source video V_S with a known text description τ_S, and generate a target video V_T using a prompt τ_T. We then record Q features from V_S, inject them into the generation of V_T, and analyze the effect of that injection. We base our main analysis on VideoCrafter2 [5] model, and later test Q-injection in other architectures for generality.

More specifically, given the video V_S, we add noise at a level corresponding to a noisy step t, yielding a noisy latent z_t. We then perform a single DDPM denoising step (with a 50-step schedule) and record the Q features from all *spatial* self-attention layers of the diffusion model. This is repeated 20 times for various noise levels, resulting in a sequence of 20 Q tensors $(Q_S(50), Q_S(49), ..., Q_S(30)$, corresponding to DDPM steps $t = 1000, 980, ..., 600)$. Finally, we generate a new video V_T with prompt τ_T, while injecting $Q_S(t)$ tensors at the first k DDPM denoising steps. We vary the amount of steps receiving Q-injection: From none (0%) up to 40% of DDPM steps.

To understand the effect of Q-injection, we measure similarity between source and target videos in two aspects: Identity Leakage and Motion Fidelity. Identity Leakage measures mean DINO similarity between the frames of source V_S and target V_T videos. Motion Fidelity [30] measures cross-correlation between point tracks in source and target videos. Error bars are standard-error-of-the-mean (S.E.M) to show the significance of our findings.

Fig. 2 shows our results. We consider two Q-injection setups: one where the source and target prompts share the same subject (purple), and one where they differ (green). We highlight three key differences in how Q injection behaves in T2V models versus T2I models.

First, Fig. 2 (bottom-left), shows that Motion Fidelity increases with Q-injection duration, reaching high similarity at 40% injection. We find that, unlike images–where structure is established early in the denoising process–videos require significantly more steps to set the motion structure. Second, more surprising is the top-left panel. It reveals that identity similarity also increases with the duration of injecting Q. This suggests that video generation models also encode identity information into the Q vectors–an intriguing shift from its traditionally assumed role in T2I models. Third, we observe an interesting phenomenon: when τ_S and τ_T use the same subject, the target video often features a subject with an appearance identical to that of the source video while maintaining the background specified in τ_T. This identity "leakage" is significantly less pronounced when τ_T and τ_S feature different subjects. Qualitatively, in Fig. 2 (right), we present a source video of "A horse gallop-

Figure 4. **Motion Transfer (VideoCrafter2).** Source (top) and target (bottom) frames for each pair. Q-injection transfers camera motion (top), non-rigid movement (middle), and combined motions (bottom). Videos and more examples in suppl.

ing in the savanna" with a distinct white spot on its back. Notice that as the number of injection steps increases, the target horse (purple) becomes more similar to the source horse while preserving the cloud background. However, when τ_T describes a giraffe, its shape and motion become more similar to those of the source horse, but its appearance remains a giraffe. This suggests Q-injection transfers both appearance and motion when subjects match, while for different subjects, the transfer primarily affects motion.

We also examine Q-injection with extended attention – a widely used technique that enables cross-video feature sharing for consistency [3, 26]. Extended attention allows frames to share K,V features across different videos, promoting visual consistency but at the cost of motion quality. As shown in Fig. 3, when generating a batch of 3 videos, this feature sharing induces "motion freeze," where natural motion patterns are suppressed due to the synchronization of features across videos. To recover motion fidelity, significantly long Q-injection periods are required. Moreover, this extended setting increases identity leakage, making it difficult to disentangle motion from identity.

Generality. We verified our findings across architectures: WAN 2.1 [27] (DiT), T2V-Turbo-V2 [14] (8-step sampling), and LTX-Video [11] (fast DiT). All models showed identity leakage for same-subject prompts and primarily transferred motion for different-subject prompts (see appendix Fig. A.2).

3. Application 1: Motion transfer

Our first application is motion transfer: given a video V_S with specific motion patterns, we generate a new video V_T following the same motion. Our approach, "*Motion by Queries*" follows the experiment described in section 2. It extracts Q-features from V_S by denoising to various timesteps ($t = 1000, 980, \ldots, 600$), obtaining $[Q_S(50), \ldots, Q_S(30)]$, then injects these during generation of V_T with prompt τ_T. For WAN 2.1, we extracted the Q-features from a single low-noise timestep and injected them into all higher-noise steps, inspired by [16] (see appendix for details).

3.1. Experiments

We evaluate quantitatively on the Motion-Transfer benchmark of [28] (66 video prompts, 22 source videos from DAVIS [22] and WebVID [2]), using VideoCrafter2 (VC2) as base model.

Baselines: (1) DMT [30]: test-time optimization; **(2) MI** [28]: fine-tuning for motion-specific embeddings; **(3) MC** [16]: zero-shot using temporal attention maps. MI*, MC* denote our reproductions. We include VMC [13] and MD [32] results reported by [28].

Evaluation Metrics: We measure Motion Fidelity (M. Fidel.), point track correlation between videos (Sec. 2); Temporal Flicker (T. Flick.), consecutive frame MAE [12]; Text Similarity, CLIP-Text score [23]; Temporal Consistency (Temp. C.), consecutive frame CLIP similarity [12]; Identity Leak (Id. Leak), DINO similarity between source/target frames; and VBench quality metrics (Aesth., Smooth., Bk Cons.). We also compare runtimes on NVIDIA H100 (576×320), breaking them into: Invers. (inversion/feature recording), Optim. (optimization/tuning), Infer (generation), Sum (total runtime), and Overhead as the ratio of "Sum" to inference in the base model.

Quantitative Results: Our optimization-free approach achieves competitive performance (Tables 1, 2): lower identity leakage than MI* (38.6 vs. 43.7), comparable text/temporal consistency, and motion fidelity of 91.5. While MC has better identity separation (24.2), our method shows superior temporal stability (T. Flick: 95.1 vs. 86.0). Crucially, we achieve ×1.2 overhead vs. base VC2, significantly faster than MC (×12), MI (×23), and DMT (×45).

Qualitative Results: Figure 4 and results in supplemental, show that Motion by Queries successfully transfers both camera and non-rigid object motion. Comparisons with baselines (Fig. A.1, Suppl.) show our method produces faithful motion, though sometimes with lower magnitude than MI/DMT. MC shows temporal instability and MI exhibits identity leakage, confirming quantitative findings. Figure 5 demonstrates results with WAN 2.1 DiT, which prioritizes physical plausibility over pixel-exact replication, particularly visible in non-rigid motion scenarios.

4. Application 2: Consistent multi-shot video generation

Multi-shot video generation requires maintaining character consistency across shots while preserving natural motion – a challenge since current models excel at single clips but

	M. Fidel. ↑	T. Flick. ↑	Text ↑	Temp. C. ↑	Id. Leak ↓	Aesth. ↑	Smooth. ↑	Bk Cons. ↑
DMT	78.8	-	28.8	93.6	-	-	-	-
VMC	93.7	-	27.1	94.6	-	-	-	-
MD	93.9	-	30.4	93.3	-	-	-	-
MI	95.5	-	31.1	93.5	-	-	-	-
MI*	**97.0 ± 0.4**	92.2 ± 0.4	29.2 ± 0.6	96.8 ± 0.2	43.7 ± 1.5	52.6 ± 1.2	95.6 ± 0.3	94.5 ± 0.3
MC*	95.0 ± 0.7	86.0 ± 0.6	**29.9 ± 0.4**	95.8 ± 0.3	**24.2 ± 1.1**	**55.6 ± 0.9**	93.5 ± 0.4	93.3 ± 0.4
Ours	91.6 ± 0.9	**95.1 ± 0.3**	28.8 ± 0.6	**97.0 ± 0.2**	38.6 ± 1.8	54.1 ± 1.3	**97.3 ± 0.2**	**94.9 ± 0.4**

Table 1. **Motion Transfer Metrics.** Mean ± S.E.M.

Time [sec.]	Invers.	Optim.	Infer	Sum	Overhead ↓
Z.Scope	-	-	9	9	-
DMT	260		150	410	×45
MI	9	190	9	208	×23
AnimateDiff	-	-	9	9	-
MC	0.3	0	104	104	×12
VC2	-	-	58	58	-
Ours	12	0	58	70	**×1.2**

Table 2. **Runtime Comparison.** Existing methods: ×12-45 overhead. Ours: ×1.2.

Figure 5. **Motion Transfer with WAN 2.1.** Source (top) and generated (bottom) frame pairs demonstrating Q-injection with DiT architecture. Videos and more examples in suppl.

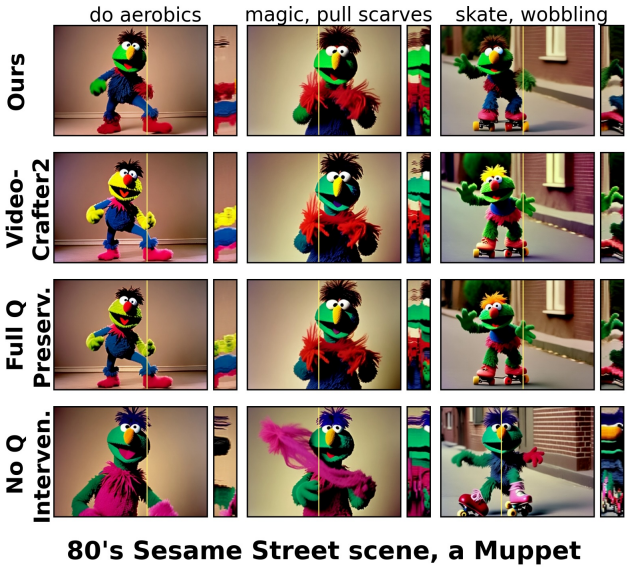

Figure 6. **Q-injection strategies for consistent video generation.** Top: Our method balances consistency and motion. Row 2: VideoCrafter2 has diverse motion, no consistency. Row 3: Full Q preservation loses character consistency because identity leaks from VideoCrafter2. Row 4: No Q intervention diminished and synchronized motion. Right: y-t slices as temporal cross-sections.

struggle with cross-shot consistency. Prior work [26] shares self-attention features for image consistency, but our analysis shows video features encode both identity and motion, causing naive extended attention to synchronize or diminish motion. We address this with a two-stage Q injection: early Q preservation sets motion structure using vanilla video Q-features, followed by Q-Flow that allows features to evolve while maintaining structure through correspondence fields .

We build on ConsiStory [26] (Appendix B.2) which shares K,V features within subject masks for consistency but reduces layout variability. ConsiStory restores diversity via Q injection from vanilla sampling. However, as shown in Fig. 3, vanilla Q injection in videos causes identity leakage. Our two-stage solution addresses this: (1) Q preservation injects vanilla Q-features early to establish motion structure; (2) Q-Flow applies correspondence fields [10] to maintain motion patterns while allowing features to evolve for better consistency. Additional details are described in Appendix B.11.

4.1. Experiments

We evaluate Q-injection strategies for consistent multi-shot generation (Fig. 6). Without Q intervention (row 4), the Muppet's identity is preserved but motion degrades: all shots show synchronized swaying, static camera in the skating shot, and frozen body with displaced legs. Vanilla Q injection (row 3) restores dynamic motion but the Muppet's colors revert to vanilla model colors, losing consistency – and demonstrating Q's dual role: restoring motion while leaking non-consistent identities. Our approach (row 1) balances both, preserving the Muppet's consistent appearance while maintaining distinct motions—centered swaying, dynamic skating with parallax camera movement. Extensive evaluations with baselines, user studies, and ablations are in Appendix B.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer, 2023. 1

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 3

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 1, 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 13

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2

[6] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9109–9137. PMLR, 2024. 11

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 7

[8] Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 13

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 13

[10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 4, 10

[11] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3

[12] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 10, 11, 13, 14

[13] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9212–9221, 2024. 3

[14] Jiachen Li, Long Qian, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design, 2024. 3, 11, 14

[15] Yumeng Li, William Beluch, Margret Keuper, Dan Zhang, and Anna Khoreva. Vstar: Generative temporal nursing for longer dynamic video synthesis. *arXiv preprint arXiv:2403.13501*, 2024. 11, 14

[16] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[17] Chao Liu and Arash Vahdat. Equivdm: Equivariant video diffusion models with temporally consistent noise, 2025. 1

[18] Timo Lüddecke and Alexander S Ecker. Prompt-based multi-modal image segmentation. *arXiv preprint arXiv:2112.10003*, 2021. 13, 16

[19] Barak Meiri, Dvir Samuel, Nir Darshan, Gal Chechik, Shai Avidan, and Rami Ben-Ari. Fixed-point inversion for text-to-image diffusion models. *arXiv preprint arXiv:2312.12540*, 2023. 7

[20] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 13, 16

[21] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23051–23061, 2023. 1

[22] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 3

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine

tuning text-to-image diffusion models for subject-driven generation. 2022. 13

[25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 16

[26] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 1, 2, 3, 4, 9, 11, 12, 13

[27] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3

[28] Luozhou Wang, Ziyang Mai, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*, 2024. 1, 2, 3

[29] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 13

[30] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 1, 2, 3

[31] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 10

[32] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024. 1, 3