Identity-Motion Trade-offs in Text-to-Video Generation

Yuval Atzmon NVIDIA Research
yatzmon@nvidia.com
Rinon Gal
Yoad Tewel
Yoni Kasten
Gal Chechik

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 and examine the challenges arising when these representations interact. Our analysis reveals 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 technique for consistent multi-shot video generation, where characters maintain identity across multiple video shots while Q-injection enhances motion fidelity. Project page: research.nvidia.com/labs/par/MotionByQueries/

1 Introduction

Video generation from text is at the forefront of generative AI. Despite progress in controlling entities in video, several major challenges remain, including generating natural, engaging motion and preserving consistent identity of entities throughout the video. These two goals often form a trade-off: It is easy to preserve consistency if motion is strongly limited, and making entities move makes it harder to enforce consistency, as the appearance of an entity changes. A major challenge lies in understanding how motion and identity are represented in various components of video generation models and how to effectively control them.

This limited understanding hinders downstream applications in video generation. For instance, while many current motion transfer approaches [34, 52, 58, 63] rely on tuning or test-time optimization, there is growing interest in developing inference-time methods, similar to how text-to-image layout transfer already operates through feature manipulation [1, 7]. 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.

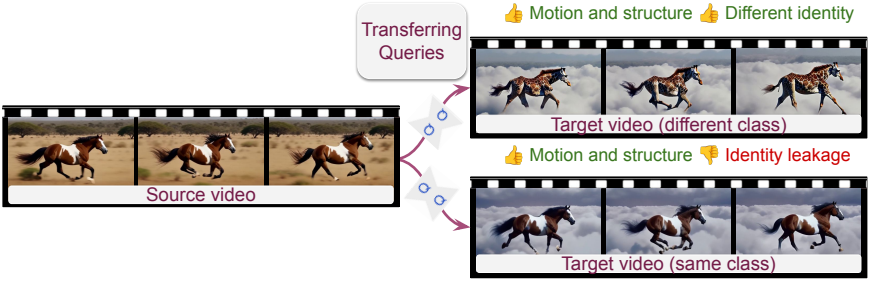

Figure 1: **(click-to-view-online)** Our analysis reveals differences in Q-injection between text-to-video and text-to-image models. One key observation is that in text-to-video models, zero-shot Q-injection can transfer structure and motion between different video shots. However, when a target video is prompted for the same subject, Q-injection suffers from identity leak.

To understand representation of motion and identity in a video model, let us first draw analogies from text-to-image (T2I) models which are better understood. Motion can be viewed as a "3D-shape" in the 3D tensor defined by a sequence of frames, so it is natural to look at representation of shape and structure in image generation models. Previous works showed that diffusion-based T2I models establish the layout of the generated image in early steps [39]. Several studies [1, 7, 49] analyzed how different model components determine this structure. They showed that self-attention queries (Q) encode structural layout information, 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, and observe that in contrast to image models, Q in video-generation models affects both motion and identity. Moreover, videos require more denoising steps than images to capture motion patterns. We use this insight to control motion in two different applications: motion transfer and consistent multi-shot video generation.

For motion transfer, we find that injecting Q features from a source video during denoising of a generated video allows transferring motion to a new video in a zero-shot manner (without extra fine-tuning or optimization). Our simple pipeline achieves generation quality close to leading methods, while being 10× more efficient than existing approaches.

For consistent multi-shot video generation, we build on insights from multi-shot image generation [49] relying on shot-to-shot extended attention. We find that Q-injection from unconstrained video generation can preserve layout and motion diversity. However, unlike images, video generation requires more Q-injection steps, causing identity leak from the unconstrained source video and compromising shot-to-shot consistency. We address this with a two-phase approach: Q-Preservation maintains motion structure using Q values from unconstrained generation, while Q-Flow instead preserves feature flow maps, avoiding leaked identity in later steps.

Our main contributions: 1) We provide a systematic analysis of Q-features in text-to-video diffusion models, revealing their dual role in encoding both motion and identity information, with effects persisting longer into the denoising process. 2) We introduce "*Motion*

by Queries", an efficient approach that enables zero-shot motion transfer in both UNeT and Diffusion-Transformer architectures. 3) We present a training-free method for consistent multi-shot video generation that balances character consistency and motion quality.

2 Related work

Motion Transfer in Video Generation: Motion transfer guides video generation using source video motion patterns, spanning camera movements and object deformations while maintaining structural integrity. Common approaches involve model adaptation (fine-tuning temporal attention [27, 43], LoRAs [63], motion embeddings [62]) or test-time optimization [58]. Concurrent noise-warped methods [6, 34] require weeks of training on 8/64 GPUs to inject motion via warped noise, and are out of the scope of our zero-shot setup. In contrast, our method enables zero-shot motion transfer via query injection without extra training or optimization. In [32], a zero-shot approach is also employed using temporal attention maps, but incurs $\times 12$ denoising overhead. Our work analyzes motion in a broader context through spatial Q features, providing principles applicable across video generation tasks and models.

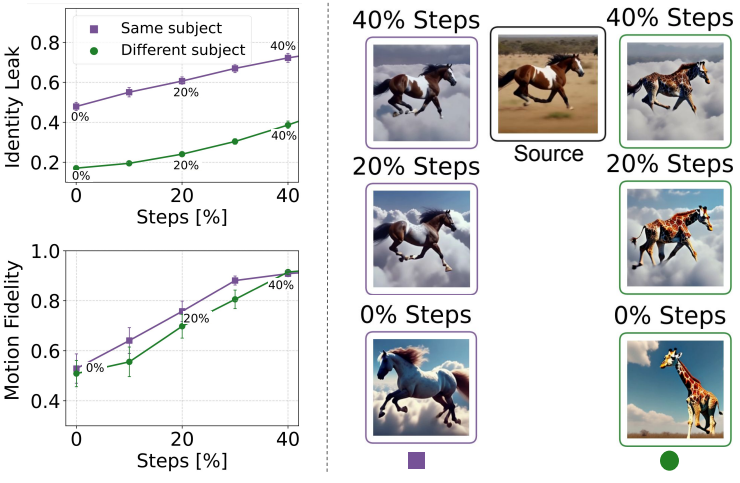

Figure 2: Same-class motion transfer suffers from identity leak (purple), worsening with increased Q-injection. Cross-class transfer (green) achieves reasonable separation at 40% injection, where motion quality is also preserved. The leftmost purple data point shows results for random same-class images (no injection). Quantitative results use motion-transfer data from [62]. For Illustration we use frames from the videos in Fig. 1.

Video Editing: Video editing modifies specific video attributes while preserving others, rather than replicating motion patterns. Bai et al. [4] also use feature injection for edit and do not address our broader Q-injection context. Zhao et al. [64] combine ControlNet and CLIP. Image-based methods [13, 35, 41, 46, 55] struggle with coherence, as they lack a temporal model. **Structure and Appearance in Text-to-Image Models:** Early work in text-to-image showed that Keys and Queries guide structure ("Where"), while Values set appearance ("What") Tewel et al. [48]. Further studies established that the layout is set early in denoising [39], and self-attention queries encode structural layout [1, 49]. We extend these findings to video, showing queries play a more complex role than previ-

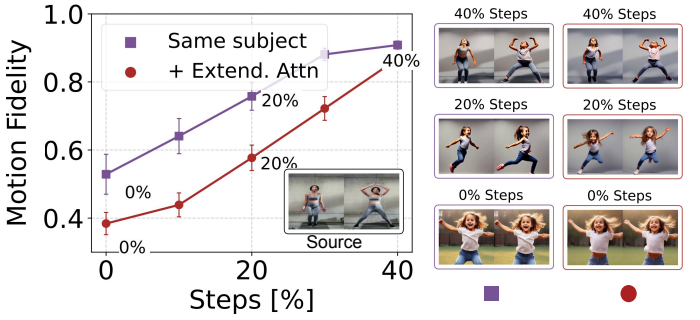

Figure 3: Motion is compromised with extended attention across video-shots. To recover the motion, longer Q-injection periods are required, which consequently increases identity leak.

ously understood - affecting both motion and identity. Additionally, we find that capturing motion patterns in videos requires more denoising steps compared to images, highlighting fundamental differences between these domains. **Extended Attention Sharing:** When using text-to-image models to generate [9, 29, 55] or modify a video [21], an extended self-attention block [55] is often employed to share keys and values across different frames, enabling them to draw visual appearances from each other and enhance consistency. Beyond cross-frame consistency, it has been used to inject consistent identities from a source image to video [10, 24, 50, 57], maintain appearance in layout editing [3, 7], combine appearances [1], for personalization [20, 51] and style transfer [23]. **Consistent generation** aims to maintain consistent subjects across outputs from a generative model, and has mainly studied in text-to-image tasks. A common approach is to use personalization [18, 44] to promote consistency, either through inpainting with a personalized model [26], personalized LoRA models [45], or training LoRAs for large, semi-consistent clusters [2]. Alternatively, encoders can inject identity at inference [19, 54, 59], but require large pre-training and struggle with new domains. Storyboard-based methods [16, 33], and video personalization methods [28, 56, 60, 62] face similar limits. Recent work [15, 49] achieves character consistency without personalization or tuning, using feature sharing across images.

3 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 [11] 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

Figure 4: (click-to-view-online) **Qualitative Results, Motion Transfer (VideoCrafter2).** Each pair of rows are frames from source (pair-top) real video and target (pair-bottom) generated video. Transferring Motion by Queries allows to use source videos to inject camera motion (top), non-rigid movement (middle), and combinations of movements (bottom). See the supplemental for videos and more examples.

[58] measures cross-correlation between point tracks in source and target videos:

$$\frac{1}{m}\sum_{\tilde{p}\in\tilde{\mathcal{P}}}\max_{p\in\mathcal{P}}\textbf{corr}(p,\tilde{p})+\frac{1}{n}\sum_{p\in\mathcal{P}}\max_{\tilde{p}\in\tilde{\mathcal{P}}}\textbf{corr}(p,\tilde{p})$$

where $\mathcal{P}=\{p_1,\ldots,p_n\},\tilde{\mathcal{P}}=\{\tilde{p}_1,\ldots,\tilde{p}_m\}$ are 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 galloping 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 specified in the prompt. On the other hand, 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 a clear distinction in Q-injection behavior: it transfers appearance and motion when the subjects match, while for different subjects, the transfer primarily affects motion characteristics.

Finally, we examine the effect of Q-injection when the target consists of multiple videos

sharing generation features. Feature sharing within a batch is widely used in T2I generation, aiding tasks like consistent character generation [49] and editing [7]. A common technique is the extended attention mechanism, which modifies self-attention layers to attend across elements of multiple videos in the same batch. Here, we explore the impact of Q-injection when generating a batch of 3 videos at a time and using shared Extended Attention between shots. In Fig. 2, we include the Extended Attention curve to visualize motion fidelity in this scenario. Our results show that extended attention induces a "motion freeze" effect, requiring Q to be injected for more steps to match the source video's motion. Moreover, since Q-injection in this setting needs more steps, it also introduces identity information, making it difficult to disentangle motion from identity injection.

Generality Across Architectures. We extended our core Q-injection experiment of Fig. 1 to several T2V models: (1) WAN 2.1 [51], a SoTA DiT model; (2) T2V-Turbo-V2 [30], a fast sampling model, using as few as 8 denoising steps; and (3) LTX-Video [22], a fast DiT model. As in our main findings, Q-injection led to identity leak for same-subject prompts and primarily transferred motion for different-subject prompts. We also verified that without Q-injection, source and target prompts produced visually distinct outputs. Qualitative results are provided in the supplemental and Fig. 8. See Appendix for implementation details.

4 Application 1: Motion transfer

Our first application is motion transfer. It allows creators to control motion in a fine way, which cannot be achieved using text. In this problem we are given a video V_S that contain some pattern of movement, and we wish to generate a new video V_T that follows the same movement patterns. The movement can be of entities in the video, or the camera or both.

Our approach, called "*Motion by Queries*" follows the experiment described in section 2. First, we extract a series of Q-features from V_S, by denoising the model to various timesteps ($t = 1000, 980, \ldots, 600$), obtaining $[Q_S(50), \ldots, Q_S(30)]$. Then, we inject those Q features during the generation of the target video V_T, this time with the new prompt τ_T (see Alg. 1). For WAN 2.1, we extracted Q-features from a single low-noise timestep and injected them into all higher-noise steps during target generation, inspired by [32]. See appendix for additional implementation details.

4.1 Experiments

We evaluate Motion by Queries qualitatively and quantitatively using a Motion-Transfer benchmark [52], comprising 66 video prompts, guided by 22 source videos. Videos were sourced from DAVIS [40], WebVID [5], and online resources, representing diverse scenes, objects, and motion. For the base text-to-video model, we use VideoCrafter2 (VC2).

Baselines: We compare our approach with recent methods: **(1) Diffusion Motion Transfer (DMT)**, a test-time optimization approach [58], a **(2) Motion-Inversion (MI)** [52], a fine-tuning method that learns motion-specific temporal embeddings from a source video, and **(3) MotionClone (MC)** [32], a zero-shot approach where motion is guided by temporal transformer self-attention maps extracted when U-Net processes source videos. MI*, MC* indicate results we reproduced from their online code. Except for one case in the supplemental, all qualitative comparisons use videos from MI and DMT project pages. We also adopt from [52] the quantitative results of Video Motion Customization (VMC) [27] and Motion-Director (MD) [63], both are fine-tuning-based approaches.

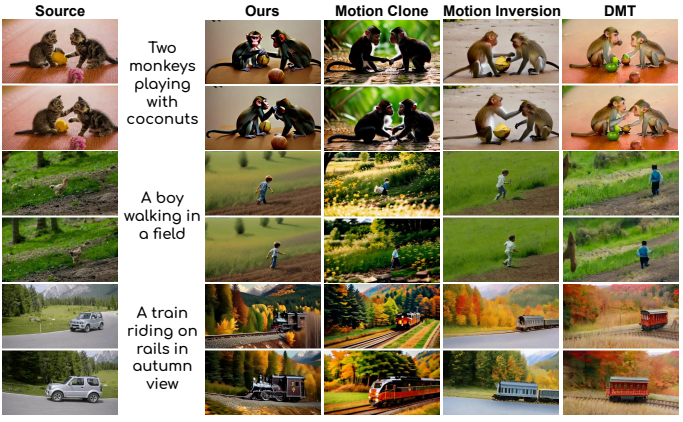

Figure 5: (click-to-view-online) **Qualitative Comparisons, Motion Transfer.** Each column shows keyframes from a different case. Boy and train videos are comparable to baselines; monkey shows a lower motion magnitude but preserves source motion pattern.

Evaluation Metrics: We measure Motion Fidelity (M. Fidel.), the point track correlation between generated and source videos, as described in Sec. 3; Temporal Flicker (T. Flick.) which quantifies flickering artifacts, using mean absolute error between consecutive frames, as in VBench [25]; Text Similarity (Text), the average CLIP-Text score between generated frames and the textual prompt [42]; Temporal Consistency (Temp. C.) which measures video coherence by averaging CLIP-Image similarity between consecutive frames [25]; Identity Leak (Id. Leak) to quantify the similarity between source and target video frames, as described in Sec. 3; Finally, Aesth., Smooth., and Bk Cons. which are VBench quality metrics for visual appeal, temporal smoothness, and background consistency.

Additionally, we compare runtimes and overhead relative to each method's base model. We break the runtime to components: "Invers." is for inversion or feature recording from source video, "Optim." for optimization or tuning, "Infer" for sampling a new video, "Sum" is the total runtime, and "Overhead" is the ratio of "Sum" to inference in the base model. Runtimes were measured on NVIDIA H100, when generating at a resolution of 576×320.

	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: **Quantitative Evaluation Metrics, Motion Transfer.** Values are mean ± S.E.M.

Quantitative Results: Our approach demonstrates competitive results across various metrics while remaining simple and optimization-free (Tables 1, 2). Specifically, our method achieves lower (better) identity leak (38.6) than MI* (43.7), with text alignment (28.8) and temporal consistency (97.0) comparable to baselines. While our motion fidelity (91.5) is moderately lower than MI* (97.0), and identity leakage (38.6) is better than MI* (43.7), both metrics remain within practical ranges as demonstrated by qualitative results, justifying our 10× efficiency gain. Compared to MC, our method attains higher identity leak (38.6 vs. 24.2, lower is better), while achieving improved temporal stability, as reflected in a higher flickering score (95.1 vs. 86.0). In terms of efficiency, our method requires only ×1.2 overhead (70s) relative to the base model (VC2), and is more efficient than MC (×12, 104s), MI (×23, 208s), and DMT (×45, 410s).

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 are ×12-45 slower than their base model. Ours adds merely ×1.2 overhead.

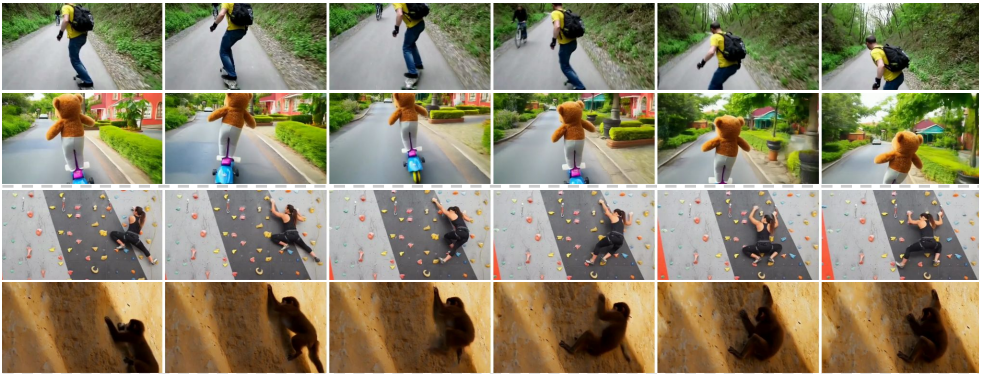

Figure 6: **(click-to-view-online) Wan 2.1, Motion Transfer.** Comparing source (top) to generated (bottom) in each pair. See supplemental for videos and more examples.

Qualitative Results: In Figure 4 and supplemental, we show that Motion by Queries guides videos generated by VC2 to follow the motion in source videos, capturing both camera motion (top), and non-rigid object movement (bottom). Figure 5 compares our results with MC, MI, and DMT: in the monkeys example, our method produces motion that resembles the source video, albeit with slightly lower motion magnitude than the baselines. For the boy and train examples, our results are comparable to those of MI and DMT. Supplemental comparisons further reveal cases where MC exhibits temporal instability and MI identity leak, as seen in the quantitative results. They also provide examples illustrating custom camera motion and hybrid object-camera motion. Figure 6 and supplemental present results with the WAN 2.1 DiT model. Compared to VC2, WAN exhibits a stronger physical prior: it tends to generate motion that is physically plausible and realistic, sometimes prioritizing physical accuracy over exact pixel-level replication. This is noticeable in non-rigid motion scenarios, such as the wall climbing example. We also find that WAN is more sensitive to the choice of initial noise, resulting in greater variability in the generated outputs.

5 Application 2: Consistent multi-shot video generation

Generating long, coherent sequences remains difficult for text-to-video models, yet this ability is key for rich storytelling and creative video applications. A practical approach is to generate several short video shots with the same characters, as in cinematic videos. The main challenge is keeping character consistency across shots, since current models excel at single clips but struggle with cross-shot generation.

Prior work [15, 49] tackled character consistency in text-to-image by sharing self-attention features. However, as our analysis shows, video features encode both identity and motion, making it challenging to apply these methods to video, where naive extended attention can cause synchronized or diminished motion. To address this, we propose a method tailored for

Algorithm 1 Q-Injection for Motion Transfer

Require: Source V_S, prompts τ_S, τ_T, max injection step k

 # Step 1: Extract Q-features
 # Q_S^l = queries from layer l of source
if WAN 2.1 **then**
  Add noise to V_S at step $i = 45$
  Store queries $\{Q_S^l\}$ from layers 20-30
else if VideoCrafter2 **then**
  for $i = 0$ to $k-1$ **do**
    Add noise to V_S at step i
    Store $\{Q_S^l(i)\}$ from all spatial attn layers
 # Step 2: Generate with Q-injection
Init random noise z_T
 # Q = target queries, Q_S = source queries
for $i = 0$ to $k-1$ **do**
  Replace: $Q^l \leftarrow Q_S^l$ or $Q_S^l(i)$ *# during denoise*

for $i = k$ to 49 **do**
  Standard denoise with prompt τ_T
Output: V_T with motion from V_S

Algorithm 2 Q-Flow for Motion Fidelity

Require: For each shot:
Vanilla video V_S w/o consistency,
target prompt τ_T, $t_{pres} = 750$

Extract $\{Q_S^l(t)\}$ from V_S (see Alg. 1)
 # Q = target queries, Q_S = source queries
Init $z_T = z_S$ *# same noise used to generate V_S*
for each consistency denoising step i **do**
  if $t_i \geq t_{pres}$ **then** *# Q-Preserve*
    Replace: $Q^l \leftarrow Q_S^l(t_i)$ *# during denoise*
  else *# Q-Flow*
    for 5-frame seg. with edges f_A, f_B **do**
      for frame f in segment **do**
        # Find NN for Q tokens
        $i_A = \arg\max \cos(Q_S^l(t_i, f), Q_S^l(t_i, f_A))$
        $i_B = \arg\max \cos(Q_S^l(t_i, f), Q_S^l(t_i, f_B))$
        # Interpolate using target Q
        $d_1 = |f - f_A|$, $d_2 = |f - f_B|$
        $w = \sigma(d_2/(d_1 + d_2))$
        $Q_{fl}^l(f) = w \cdot Q^l(f_A)[i_A] +$
          $(1-w) \cdot Q^l(f_B)[i_B]$
        $\alpha = 0.25 + t_i/1000$ *# Blend*
        Replace: $Q^l \leftarrow \alpha Q_{fl}^l + (1-\alpha) Q^l$
Output: V_T with preserved motion

consistent video generation, combining extended attention for consistency [49] with a two-step Q-injection process. Early Q preservation sets motion structure, followed by flow-based Q-injection that lets Q values evolve for better consistency while maintaining structure.

We build on ConsiStory [49] (Appendix, Sec. B.2) that applies extended attention within a subject mask to share K, V self-attention features, but reduces layout variability. To restore diversity, ConsiStory injects Q features from a vanilla, non-consistent sampling step. While this works for images, injecting Q features from vanilla videos preserves motion diversity but causes identity leak from those videos, breaking character consistency (Fig. 2).

Our two-stage solution—Q preservation followed by Q Flow—addresses this. In early diffusion steps, which primarily control motion and layout, we inject the Q-features from vanilla videos generated with the same prompts and noise but without extended attention. Then, in the subsequent *Q-Flow* phase we apply a relaxed Q-injection that preserves the flow of Q-features rather than injecting the features directly. This technique is inspired by TokenFlow [21]. Specifically, for each vanilla video, we first compute the nearest-neighbor correspondence field on its Q-features, defining a Q-feature flow to maintain in the generated consistent video. We then inject this correspondence into our generated video (see Algorithm 2 and Appendix B.11 for details).

5.1 Experiments

We investigated the effects of self-attention query (Q) tokens on motion and identity for consistent video generation described above. Further extensive evaluations, including baselines,

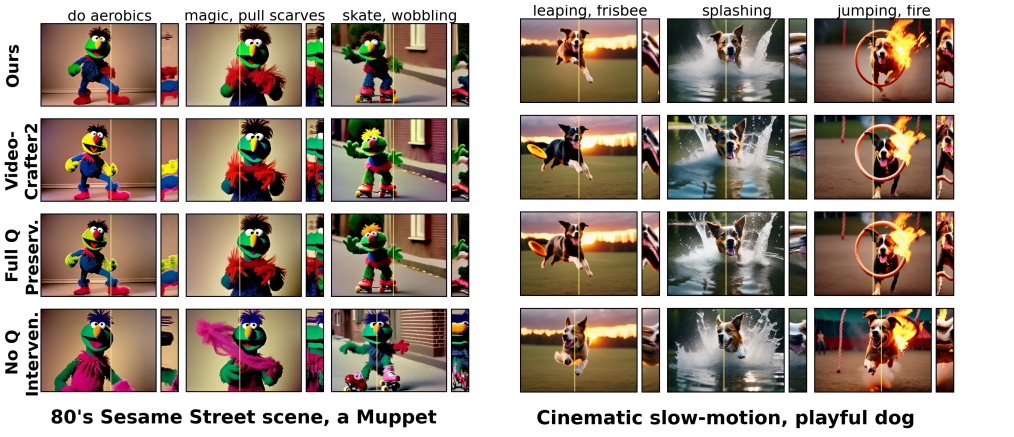

Figure 7: (click-to-view-online) **Q token strategies for consistent video** (section 5.1)**.** "Ours" (top) balances identity and motion. VC2 (second) offers diverse motion but lacks consistency. "Full Q Preservation" (third) keeps motion, but loses identity since it leaks from VC2. "No Q Intervention" (bottom) keeps identity, degrades motion. Right boxes: temporal cross-sections (y-t slices) along the yellow line, illustrating motion preservation.

user studies, and ablations, are in Appendix B. Fig. 7 shows typical generations for different Q token strategies when combined with extended self-attention.

Without Q intervention (Fig. 7, 4th row), subject identity is preserved across shots, but motion quality drops: (1) movements become synchronized, *e.g.*, the dog's head turning simultaneously in all shots. (2) motion style and pose vary less: repeatition across shots, *e.g.*, the dog's leap, the Muppet's centered swaying, the camera movement becomes static in the skating Muppet shot. (3) motion-artifacts, to reconcile motion sharing with the varying text prompts, *e.g.*, the skating Muppet's body appears frozen while its legs are displaced. In contrast, Injecting Q tokens from vanilla videos (3rd row) restores motion but loses subject identity, *e.g.* the Muppet's colors revert to those of the vanilla model. This reiterate the dual role of Q tokens: injecting vanilla Qs *restores motion*, but leaks the vanilla non-consistent identities into the multi-shot mechanism. Our approach (1st row) balances both, restoring the original motion, including nuanced details like body and face orientations, postures, specific body parts , as well as parallax camera movement.

6 Conclusion and Limitations

Our findings show that, unlike in text-to-image models, Q-injection in video models affects both motion and identity, requiring more denoising steps to establish motion patterns. This challenges the classic separation of "where" and "what" pathways [1, 39, 48], revealing a more complex interplay in video models. These insights led to two applications: zero-shot motion transfer and training-free consistent multi-shot video generation. Our method has limits: motion fidelity can drop with fast movements, and cross-class transfer may cause shape leakage from both subject and background, especially when source and target share similar scene descriptions (See Fig. 9 in Appendix). In multi-shot generation, balancing identity and motion remains hard—Q-injection can still hurt identity, and partial dropout may be needed to preserve motion (section B.13).

References

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

[2] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. *arXiv preprint arXiv:2311.10093*, 2023.

[3] Omri Avrahami, Rinon Gal, Gal Chechik, Ohad Fried, Dani Lischinski, Arash Vahdat, and Weili Nie. Diffuhaul: A training-free method for object dragging in images. *arXiv preprint arXiv:2406.01594*, 2024.

[4] Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*, 2024.

[5] 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.

[6] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, 2025.

[7] 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, October 2023.

[8] 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.

[9] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023.

[10] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*, 2023.

[11] 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.

[12] 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*, volume 235 of *Proceedings of Machine Learning Research*, pages 9109–9137. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cohen24a.html.

[13] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

[14] 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*, ICML'24. JMLR.org, 2024.

[15] Jiaojiao Fan, Haotian Xue, Qinsheng Zhang, and Yongxin Chen. Refdrop: Controllable consistency in image or video generation via reference feature guidance. *arXiv preprint arXiv:2405.17661*, 2024.

[16] Zhangyin Feng, Yuchen Ren, Xinmiao Yu, Xiaocheng Feng, Duyu Tang, Shuming Shi, and Bing Qin. Improved visual story generation with adaptive context modeling. *arXiv preprint arXiv:2305.16811*, 2023.

[17] 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. URL https://openreview.net/forum?id=DEiNSfh1k7.

[18] 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. URL https://arxiv.org/abs/2208.01618.

[19] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023.

[20] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization, 2024.

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

[22] 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.

[23] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. 2023.

[24] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

[25] 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.

[26] Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900*, 2023.

[27] 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.

[28] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. 2024.

[29] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, October 2023.

[30] 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.

[31] 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.

[32] 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. URL https://openreview.net/forum?id=aY3L65HgHJ.

[33] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, and Weidi Xie. Intelligent grimm–open-ended visual storytelling via latent diffusion models. *arXiv preprint arXiv:2306.00973*, 2023.

[34] Chao Liu and Arash Vahdat. Equivdm: Equivariant video diffusion models with temporally consistent noise, 2025. URL https://arxiv.org/abs/2504.09789.

[35] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024.

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

[37] 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.

[38] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

[39] 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.

[40] 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.

[41] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[42] 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.

[43] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 332–349. Springer, 2024.

[44] 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.

[45] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2023.

[46] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *Asian Conference on Machine Learning*, pages 1215–1230. PMLR, 2024.

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

[48] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[49] 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.

[50] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. *arXiv preprint arXiv:2311.18830*, 2023.

[51] 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.

[52] 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.

[53] 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. URL https://openreview.net/forum?id=MLBdiWu4Fw.

[54] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15943–15953, October 2023.

[55] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023.

[56] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *NeurIPS*, 2024.

[57] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. 2023.

[58] 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.

[59] 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.

[60] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. 2025.

[61] Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6786–6795, 2024.

[62] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magic mirror: Id-preserved video generation in video diffusion transformers. *arXiv preprint arXiv:2501.03931*, 2025.

[63] 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.

[64] Yuyang Zhao, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Make-a-protagonist: Generic video editing with an ensemble of experts. *arXiv preprint arXiv:2305.08850*, 2023.