Controlled video generation without finetuning

Zixuan Lin, Qingquan Bao
Philadelphia, PA
University of Pennsylvania

Abstract
Generating spatially and temporally consistent action-specific videos, such as golf swing sequences, using pre-trained diffusion models without domain-specific fine-tuning is a significant challenge in video synthesis. This project explores novel methodologies for zero-shot video generation by leveraging latent space manipulations in Stable Diffusion pipelines. Our approach demonstrates the critical role of noise progression in achieving frame consistency, highlights the limitations of naive operations in high-dimensional latent space, and introduces the potential of latent differences as a feature to guide temporal dynamics. Additionally, we propose a novel framework for action-specific video generation that utilizes latent semantic shifts derived from aligned video data, achieving scalability and eliminating the need for fine-tuning. Our findings surpass initial expectations, providing a robust foundation for future research in fine-tuning-free video generation and its applications. Codes are available at `https://github.com/QingquanBao/Controlled-video-generation-without-finetuning`.

1 Introduction

Generating spatially and temporally consistent videos remains a core challenge in computer vision, particularly for action-specific tasks such as modeling a golf swing. Achieving coherence across frames while preserving semantic details of dynamic actions and backgrounds is crucial for applications in sports analysis, virtual reality, and video synthesis.

Traditional methods often depend on computationally intensive fine-tuning of generative models on domain-specific datasets. While effective, these approaches limit scalability and adaptability to diverse scenarios. This project aims to overcome these limitations by developing a zero-shot framework for video generation that leverages pre-trained diffusion models and pose estimation techniques without domain-specific fine-tuning.

Key challenges include: (1) achieving temporal consistency across frames while retaining visual quality, (2) maintaining semantic integrity of complex actions such as human poses and object interactions, and (3) ensuring computational efficiency by avoiding additional training.

To address these challenges, we explored innovative strategies, including:

- **Noise manipulation:** Demonstrating the importance of noise progression and fixing noise seeds to enhance temporal coherence.
- **Latent space exploration:** Investigating the limitations of naive operations such as latent blending and introducing effective alternatives like intermediate latent copying.
- **Latent differences:** Identifying the potential of latent differences as a dynamic feature to guide temporal transitions and semantic shifts.
- **Leveraging original video latents:** Integrating pre-existing motion and style cues to refine generated outputs.
- **Novel action-specific pipeline:** Proposing a framework inspired by video compression, where aligned latent differences capture generic semantic shifts for scalable video generation.

Our results demonstrate the feasibility of generating temporally consistent and semantically accurate videos without fine-tuning, achieving outcomes

that exceed expectations. By focusing on latent space dynamics and noise control, this work lays a scalable foundation for future research in fine-tuning-free video synthesis and its broader applications.

2 Related Work

The problem of pose-controlled video generation has been extensively studied, with a focus on improving spatial and temporal consistency. However, existing methods often require domain-specific fine-tuning or lack the flexibility for diverse action generation.

Ma et al. [1] propose a method for pose-controllable and text-editable video generation using a two-stage training process. The first stage aligns poses with images using the LAION-Pose dataset, while the second stage refines temporal consistency through pose-free videos. Their approach employs modified latent diffusion models with temporal and cross-frame self-attention to ensure video coherence. Although effective in generating diverse characters and backgrounds, the method exhibits temporal inconsistencies in complex scenarios. Moreover, our implementation of cross-frame attention similarly showed no improvement in temporal consistency. The method's reliance on large-scale fine-tuning and dataset preparation further limits its accessibility and scalability.

Zhong et al. [2] introduce a method for pose-controlled video generation using Stable Diffusion and ControlNet. Their two-stage approach involves selecting reference frames to preserve visual details and using latent editing with affine transformations for refining faces and hands. While this method excels in maintaining detail and identity, it restricts flexibility by requiring pre-existing reference frames for identity transfer. Additionally, it does not allow for direct manipulation of learned semantics, making it less suitable for arbitrary or generalized video generation tasks.

3 Dataset

We will be using GolfDB [3], a video database for golf swing sequencing, which contains 700 sequences of golf swings under various conditions.

4 Approach

4.1 Set Up

We use OpenPose [5] to extract poses from videos in GolfDB. We choose ControlNet with Stable-Diffusion 2.1 as our baseline.

4.2 Visualization

We used N-Cut to visualize the latents when 20 golf images are processed by the stable-diffusion 2.1 pipeline. Since ControlNet is trained with stable-diffusion U-Net frozen, the insights from the visualization should carry over to ControlNet.

See Appendix for visualization.

By observing the visualization, we find that layers in the middle of downscaling and upscaling phase of U-net is more suitable for investigation and manipulation. The layers at very beginning and end are high resolution have have a lot of noise. The layers in the middle of the U-net where the spatial resolution is the lowest has very little spatial information, making them hard to interpret. We found that the latent is very sensitive to color as clothes of the same color will certainly exhibits latents of similar color when visualized using N-cut and tSNE.

4.3 Cross Frame Attention

Levon et.al. [6] introduce cross frame attention to preserve the foreground objects' semantics. In detail, the self-attention layer is modified to be a cross-attention layer where the key and value latents are replaced with the first frame's ones. It is implemented in Huggingface as an attention block call-

back. Figure 1 shows the result, and it does not present the effect of the paper showing.

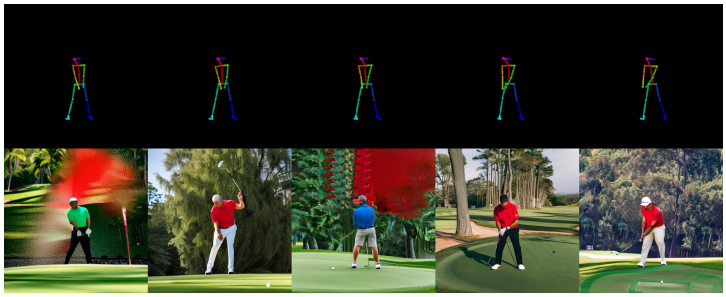

Figure 1: Video frames generated by cross-frame attention.

4.4 Copying Over Images

We attempt copying over the images from previous frames and start diffusing from there in the hope that an existing image will help guide the diffusion process. However, the results of such method is not ideal. We still see that the subsequent frames varying a lot in Figure 2. This is perhaps because the noise added in the diffusion process plays a much bigger role in determining the outcome of the process.

Figure 2: Video frames generated by diffusing from same initial latent.

Therefore, we try to investigate the effect of noise on the generation result.

4.5 Manipulation of Noise

We fix the seed of the noise generator during the diffusion process for each image. The result is surprising in Figure 3. Despite having different initial latent and control signals, the generated images are consistent, especially in the background where it is not effected by the control signal.

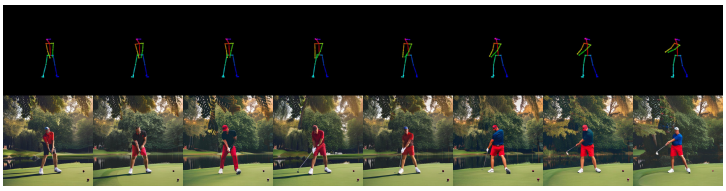

Figure 3: Video frames generated by diffusing using the same noise.

4.6 Start Diffusing from Intermediate Time Step with Copied Latent

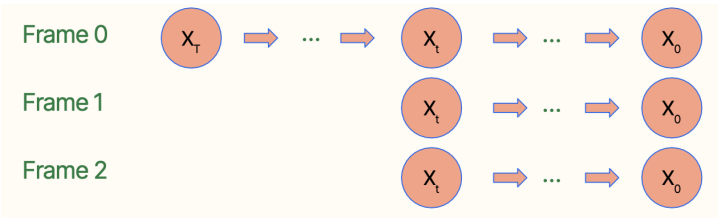

Figure 4: Illustration of the latent copying pipeline

Since noise added in the diffusion process is mostly what causes the inconsistency across frames, if we start diffusing from intermediate time step so that less noise is added during the process, it is easier to generate consistent frames. As illustrated in Figure 4, we select a time step t in the diffusion process and copy the latent when the first video frame is generated. For subsequence frames, the diffusion process start at time step t with the copied latent.

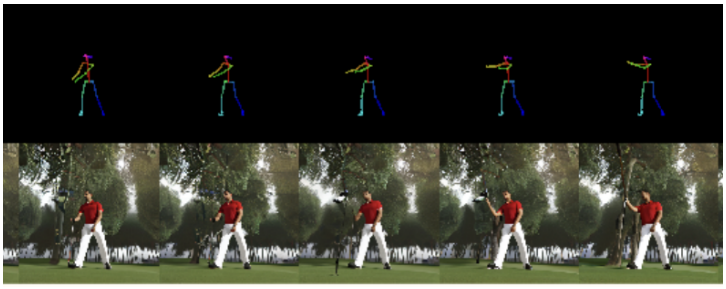

Figure 5: Effects of the latent copying pipeline at $t = 15$

We hand-picked $t = 15$ through experimentation which is shown in Figure 5. Earlier time steps do not ensure enough consistency while latter time steps do not leave enough room for ControlNet to control the postures naturally.

We also investigate the effects of blending through weighted sum of two latents of different frames. Essentially, it is a softer version of directly copying.

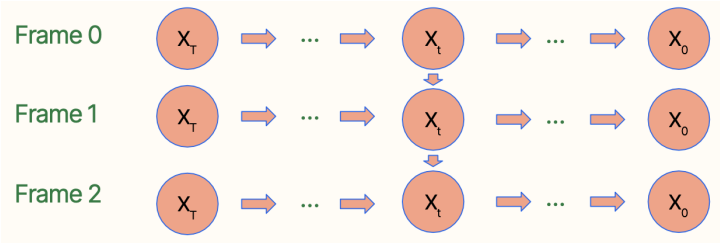

Figure 6: Illustration of the latent blending pipeline

As shown in Figure 6, each frame is still diffused from the start with different noise but latent of the first frame is blended with the latent of the frame being generated. The idea is that potentially a weaker control will allow for more natural looking image. However, the effects are not desirable.

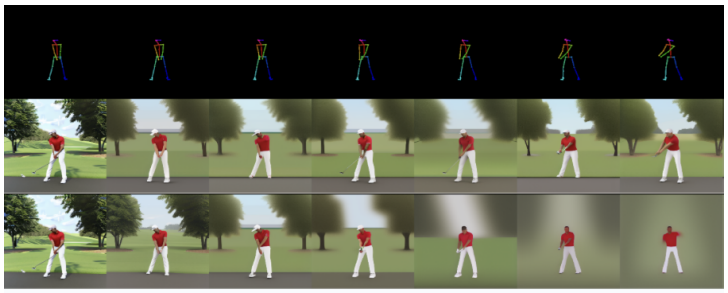

Figure 7: Effects of the latent blending. Middle: blend with latent of the first frame. Bottom: blend with latent of the previous frame.

We can see in Figure 7 that the background gets blurry if we blend two latents together. The reason is that latents are high dimensional vector lying on the manifold encoding semantic information and weighted sum of latents might not preserve the semantic meaning. Compared to the background, the human is relatively clear because the latents are more aligned at these pixels in that they most likely are related to body parts or clothes, so semantics are preserved after the operation.

4.7 Leveraging the Original Video

So far, the best result we have observed comes from directly copying the latent at intermediate diffusing time step. The background achieves satisfying level of consistency. However, the pose of the human does not conform to the action of interest. For example, the golf club is missing. This is the limitation of stable diffusion trained on generic image data as it does not have deep knowledge about the field of golf. One way to improve the generation quality would be fine-tuning on images of people performing golf swings. However, that is against the intention of the project which is no training. Despite the recent advances on training technique such as LoRA, fine-tuning a model like stable diffusion is still time consuming, let alone the effort for collecting quality data.

The skeleton used as control signals by ControlNet comes from an existing video, so we might as well harness some of the power embedded in the original video.

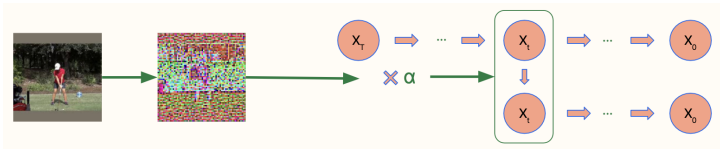

Figure 8: Illustration of the original latent injecting pipeline

We use the VAE encoder of Stable Diffusion to encode the frames of the original video. During the generation process, at a selected frame, we fuse the original latent of matching frames with the latent generated by weighted sum as shown in Figure 8.

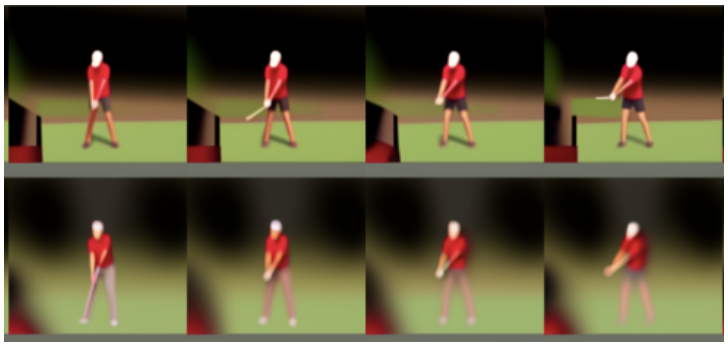

Figure 9: Effects of the original latent injecting pipeline. Top: high α. Bottom: median α.

From the Figure 9, we can observe the effects of injection. If the weight for the original latent is high (around 0.9), then the generated sequence is very similar to the original video and leave little room for control of styles. If the weight is low (below 0.4), then there is little trace of the original image and the poses look similar to the results from previous methods. If the weight is somewhat in the middle, then the images become blurry for the same reasons mentioned above.

Since the blurring effect is less significant on the human part, we try to inject the noise only to the human so that the background can be stylishly generated.

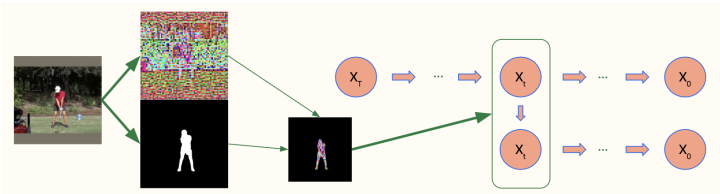

Figure 10: Illustration of the original latent injecting pipeline with masking

Masks are extracted from the original video using SAM-2 [7]. At masked pixels of the human, latents of the original video frames are fully covers the generated latents. For other pixels, the latents of the first frame are copied over like the method mentioned previously. The diagram is in Figure 10.

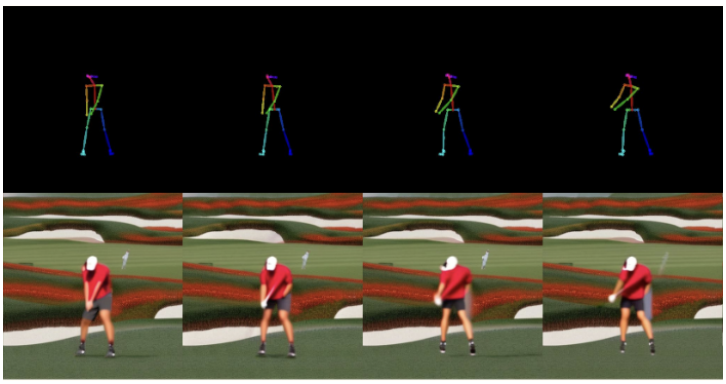

Figure 11: Effects of the original latent injecting pipeline with masking

The result in Figure 11 seems to be decent as the background is as consistent as we previously observed and the pose of the human is reasonable for a golf swing. Since the mask could not capture the golf club, the club is still missing from the generated images.

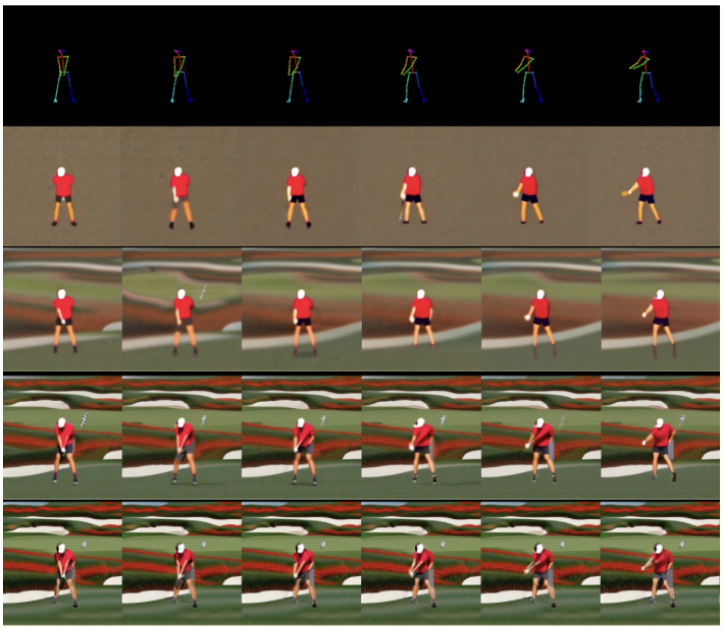

Figure 12: Comparison of injecting at different time steps

By trying out different time steps for injection in Figure 12, we found that when injecting at earlier time steps, the final outcome becomes a low definition human in a background with little texture. Our reasoning for this is that during the denoising pro-

cess, a human with a noisy background is not something that shows up in the training samples, therefore, the denoising prediction is less effective. When the U-Net of stable diffusion observes the noisy latent with a clear human in the middle, it determines that the image does not have a lot of noise, therefore predicts a noise with small magnitude. As a result, the final background is just a smoothed Gaussian noise. On the other hand, if we inject the masked latent at a latter time step, then there is not enough steps for the model to effectively blend the generated latent and original latent, creating artifacts in the outcome. Overall, it is difficult to find a balance between high definition image and low artifact image.

In previous methods, copying over the latent effectively discards the efforts made by earlier iterations of diffusion, so we start diffusing directly from a selected time step. However, in the case of masked injection, since we do not directly modify the part out side of the mask, the operation can be applied at multiple time steps repeatedly and the results will be different than if it is only applied once.

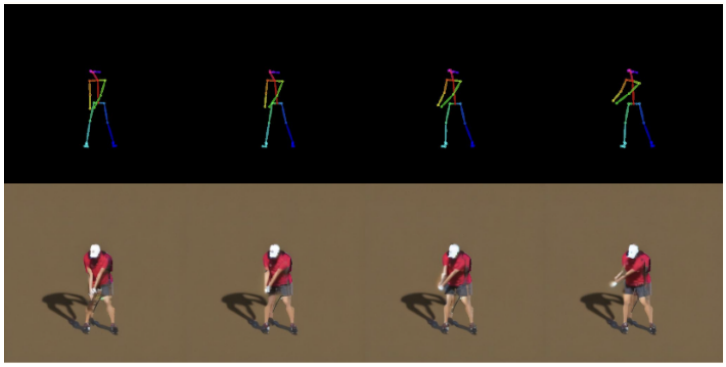

Figure 13: Effects of injecting at every single time step

We tried to apply the operation at every single step of the denoising process in Figure 13 and it is interesting to see the denoiser added a shadow for the person. Other than the shadow, the brown background and high definition human is expected.

4.8 Abandon ControlNet, Control with Latent Changes

After the all the attempt to work with ControlNet, the results have been decent but not satisfying. In addition, all the work seems basic and lacks rooms for growth. That is why the methods seem disjointed from each other and lack depth.

We aspire to come up with novel and original ideas that have potential to evolve. To this end, we abandon the framework of ControlNet and attempt to control the video generation through the changes of latents in the original video through time.

The inspiration comes from video compression. Video compression and streaming rely on breaking the video into frames, where some frames are encoded as reference frames and others as predictive frames (P-frames) or bidirectional frames (B-frames). Reference frames store complete image data, while P-frames and B-frames encode only the differences from preceding or surrounding reference frames to save bandwidth. During playback, the decoder reconstructs each frame by combining the reference frame with the stored differences, allowing efficient compression while maintaining quality.

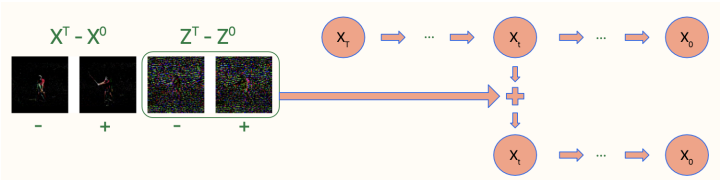

Figure 14: Pipeline of controlling with latent changes

In the new method, we obtain the latents of the original video frames and subtract the each subsequent frames from the initial reference frame. By doing so, we have the change of the latents across time. Then, the change of latents is applied to the generated latents of all but the first frame and the diffusion process begins from the modified latents. In this pipeline, skeleton control is used to generate the first frame so that the human is aligned with the

latents of the original video and no skeleton control is used in subsequent frames.

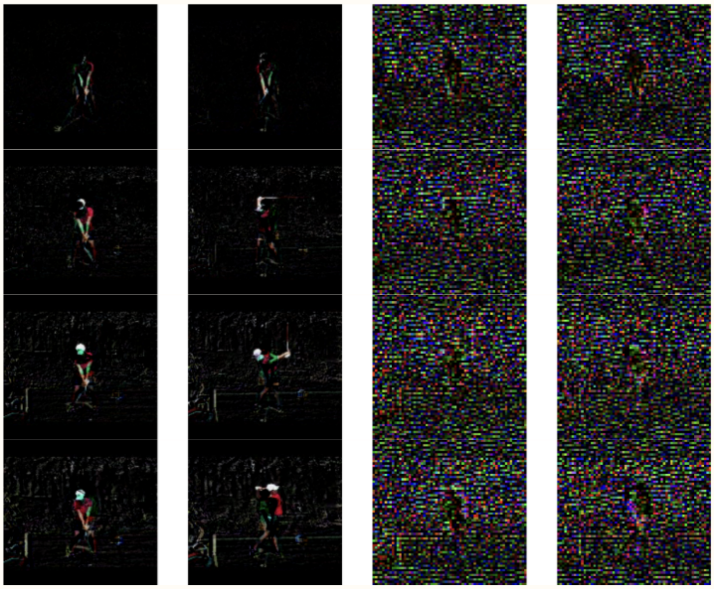

Figure 15: Even if the RGB images are very clean, the latents are still very noisy.

An interesting finding is that even though we have a video from a stationary view point, meaning the background is mostly unchanged, we still see lots of variation in the difference between latents of the background. The reason is unclear to us.

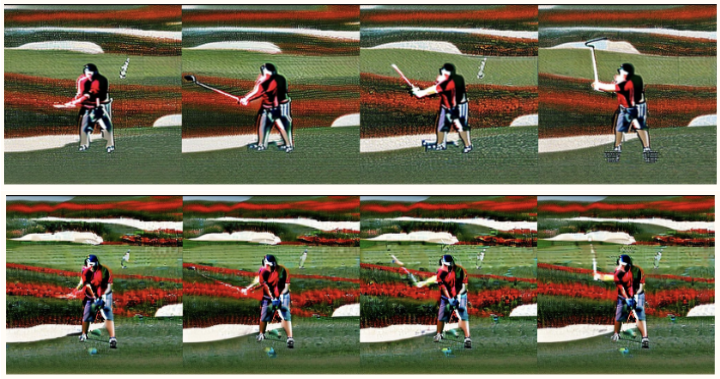

Figure 16: Effects of controlling with latent changes. Top: Control at early time step. Bottom: Control at latter time step.

We can see that even though we have lost the control signals from skeletons, the human in the generated video exhibits desired poses. Additionally, the golf club finally appears, even though the details are still missing, suggesting that the latents contains information of a stick like object but does not encode the idea of a gold club. The choice of time step to start the diffusion influences the results of generation. Starting at earlier time steps, may lead to a significant deviation of style from the first frame as well as inconsistent style among subsequent frames. Starting at the latter time steps leave fewer iterations for the denoiser to fix the artifacts introduced by the addition operation.

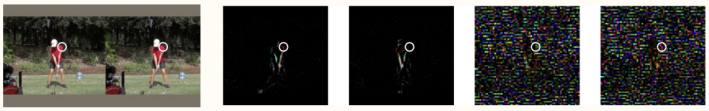

Figure 17: RGB, RGB differences, latent differences

Each pixel of latent carries semantic information at the location, and the difference between two latent pixels indicates the change of semantic information. Take the region in the white circle in the figure above as an example. We can see that the golfer's shoulder moved across the frames and the content of changed for some pixels in the region. In the RGB space, it changes from red to dark green, while in latent space, it could signify a change from the shoulder of a human wearing red shirt to the woods in the background.

5 Future Work

As briefly discussed in previous sections, difference between two pixels in the latent space captures the shift in semantic meaning. While adding two latents together may create latents without proper semantic meaning, resulting in blurry images, adding a difference to a latent will create meaningful new latent so long as the difference is extracted by subtracting from images with semantics aligned with the old latent.

The differences extracted from pair of images contains semantic shift specific to the pair and a large portion of them may not be meaningful in the

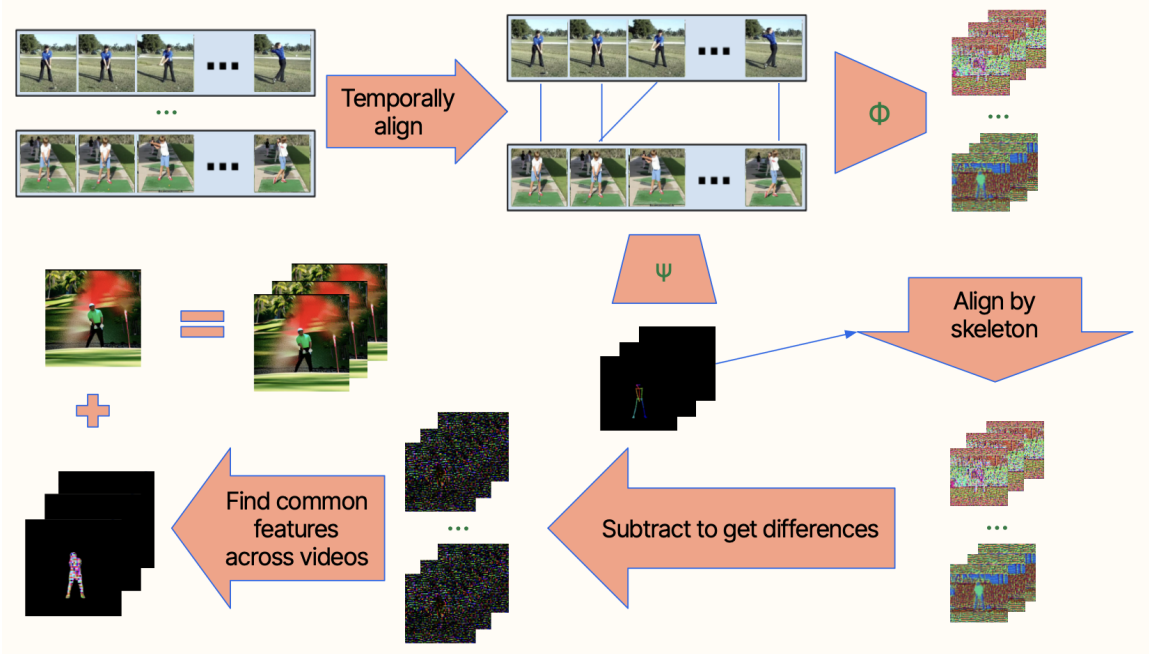

Figure 18: From a set of videos of the same action, we first align the videos temporally, then use VAE encoder to extract the latents of the frames. The latents are further aligned by body proportions extracted from skeletons. Difference between frames are computed for each video and are then used to find the generic shift of semantics across time for the action. The computed generic shift can be applied to any reference image with scaling to generate videos with the specific action while preserving the style.

general setting. Same can be said for difference extracted from a sequence of temporally related images. Hence, we see artifacts in the results shown in previous sections. Nevertheless, clean and meaningful shifts in semantics can be extracted with a sufficiently diverse set of data.

Therefore, we propose a novel framework for action specific video generation that harness the power of stable diffusion but does not require without the need for domain or task specific fine-tuning.

5.1 Novel Video Generation Pipeline

Figure 18 illustrates the full pipeline that has the potential for video generation given a diverse set of video sample of the action. After temporal and spatial alignment, the trends that is present across the latent changes extracted from all videos will embody the semantic shifts at the corresponding pixels. This semantic shift will not include information specific to any particular data sample, such as color of the shirts. Instead, it will represent a change on the level fundamental to the action of interest, such as from golf club head to background.

All of the processes denoted by arrows and trapezoids in the diagram requires no training. Off-the-shelf models and algorithms can be employed for this purpose. Temporal alignment can be computed from TCC [8]. Pose can be extracted from Open-pose [5]. Trends in the latents can be mined naively through averaging or other more sophisticated unsupervised learning methods. All of these operations are possible on modern PC and should take at most a few minutes even with hundreds of video samples.

5.2 Scaling Potential

Such method also have potential for text prompted generation. Captioned videos can be first encoded into sequences of latents then sequences of latent

changes. Compared sequences of latents, sequences of changes contains much less noise and the distribution is much more condensed. Just like the denoiser in diffusion learns to predict the noise instead of the image, learning to predict the changes given a reference frame is potentially easier and scalable than directly predicting the next frame.

6 Conclusion

In this project, we explored the challenges and opportunities in generating spatially and temporally consistent videos using pre-trained diffusion models without domain-specific fine-tuning. By systematically experimenting with noise manipulation, latent blending, and latent difference analysis, we uncovered several valuable insights:

1. **Noise Progression and Early-Stage Consistency:** The role of noise in diffusion models is pivotal for achieving temporal consistency. Fixing noise seeds or copying latents during early diffusion stages demonstrated strong improvements in background consistency across frames. These findings highlight that noise progression and its interaction with the diffusion process significantly influence video coherence. This reinforces the importance of noise control as a foundational technique, even if it appears heuristic.

2. **Latent Space as a High-Dimensional Manifold:** The experiments showed that naive operations such as blending latents from different frames or combining generated latents with raw image features often lead to semantic misalignment, resulting in blurred or inconsistent outputs. This is due to the high-dimensional nature of the latent space, where arithmetic operations like addition fail to preserve semantic meaning. Localized manipulations or carefully crafted strategies, such as masking or intermediate latent copying, proved more effective.

3. **Latent Differences as Temporal Dynamics:** Latent differences, which encode semantic transitions between frames, emerged as a promising feature for representing temporal dynamics. By leveraging these differences, we demonstrated the potential for controlling video generation in a semantically meaningful manner. This approach bridges the gap between zero-shot generation and action-specific video synthesis, enabling a more interpretable and scalable pipeline.

4. **Limitations of Skeleton Control and Fine-Tuning-Free Methods:** While skeleton-based control signals effectively guide the first frame's structure, they struggle to maintain semantic fidelity in subsequent frames, particularly for domain-specific elements like a golf club. Injecting original video latents improved action fidelity but introduced challenges in balancing fine detail and artifact reduction. These results emphasize the limitations of pre-trained models in domain-specific contexts and the potential need for task-specific data curation or lightweight fine-tuning.

5. **Proposed Framework for Action-Specific Video Generation:** Building on the insights from latent difference analysis, we proposed a novel framework for action-specific video generation. By aligning temporal and spatial dynamics across diverse video datasets, the framework captures generic semantic shifts fundamental to the action of interest. This method offers a promising pathway for scalable video synthesis without domain-specific fine-tuning, leveraging off-the-shelf models and algorithms for alignment and trend extraction.

Overall, this project achieved a level of spatial and temporal consistency in video generation that exceeded initial expectations for a zero-shot setup. The findings underscore the feasibility of leveraging pre-trained models for video synthesis and contribute valuable methodologies for future research. Further work will focus on refining these techniques

to reduce artifacts, enhance domain-specific action fidelity, and extend the pipeline to multi-view and text-prompted video generation. By doing so, we aim to advance the capabilities of zero-shot and fine-tuning-free generative methods for diverse applications.

References

[1] Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos
https://arxiv.org/pdf/2304.01186

[2] PoseCrafter: One-Shot Personalized Video Synthesis Following Flexible Pose Control
https://arxiv.org/pdf/2405.14582

[3] GolfDB: A Video Database for Golf Swing Sequencing
https://github.com/wmcnally/GolfDB?tab=readme-ov-file

[4] Kumari, Nupur and Zhang, Bingliang and Zhang, Richard and Shechtman, Eli and Zhu, Jun-Yan. 2023, Multi-Concept Customization of Text-to-Image Diffusion, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

[5] OpenPose: Real-time multi-person keypoint detection library for body, face, hands, and foot estimation
https://github.com/CMU-Perceptual-Computing-Lab/openpose

[6] Khachatryan, Levon and Movsisyan, Andranik and Tadevosyan, Vahram and Henschel, Roberto and Wang, Zhangyang and Navasardyan, Shant and Shi, Humphrey. 2023, Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators, arXiv preprint arXiv:2303.13439

[7] Nikhila Ravi and Valentin Gabeur and Yuan-Ting Hu and Ronghang Hu and Chaitanya Ryali and Tengyu Ma and Haitham Khedr and Roman Rädle and Chloe Rolland and Laura Gustafson and Eric Mintun and Junting Pan and Kalyan Vasudev Alwala and Nicolas Carion and Chao-Yuan Wu and Ross Girshick and Piotr Dollár and Christoph Feichtenhofer. 2024, SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714

[8] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2019, Temporal Cycle-Consistency Learning. arXiv:1904.07846v1

Appendices

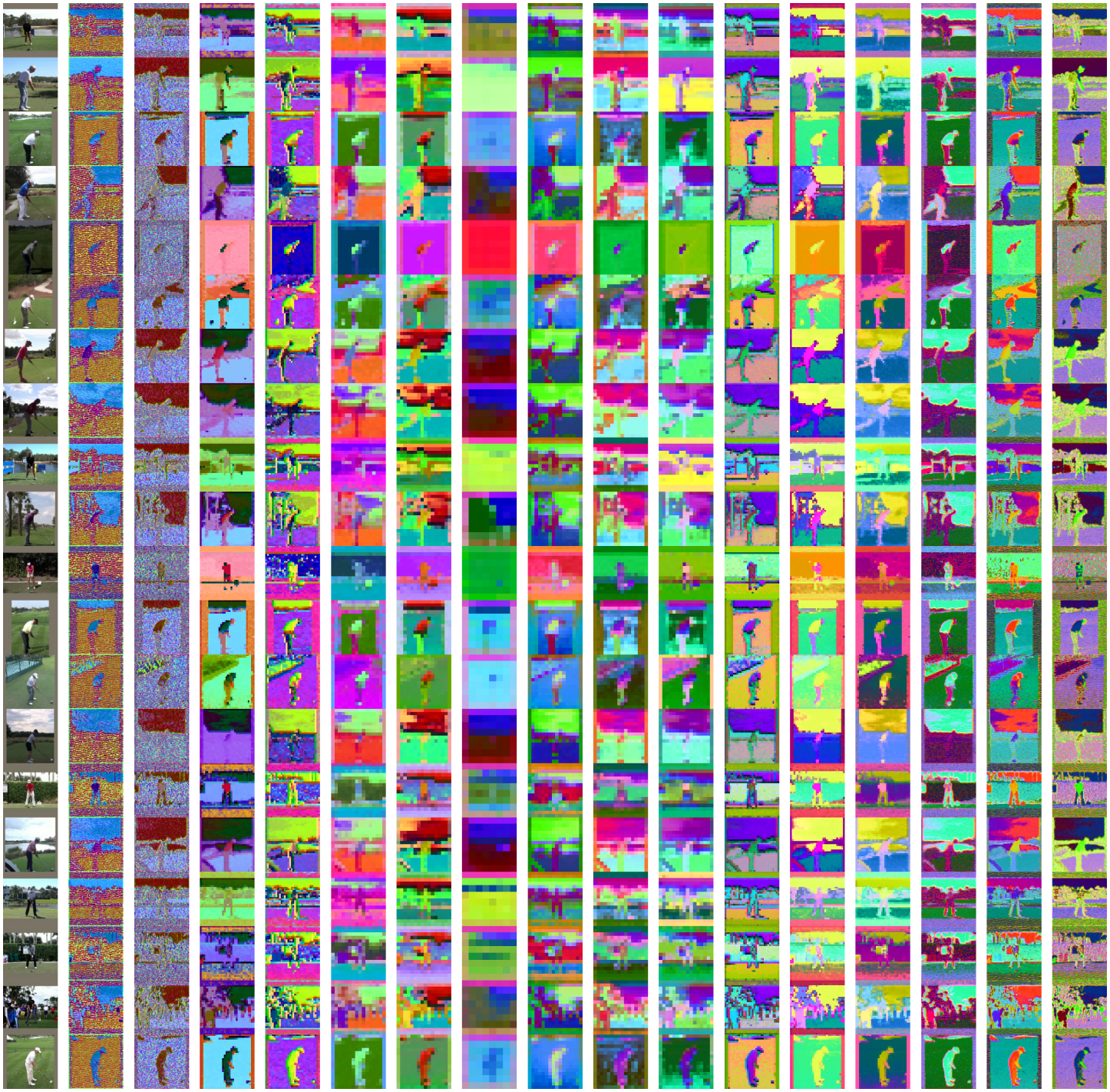

Figure 19: Attention Block