LATENTMAN : Generating Consistent Animated Characters using Image Diffusion Models

Abdelrahman Eldesokey
KAUST, Saudi Arabia
abdelrahman.eldesokey@kaust.edu.sa

Peter Wonka
KAUST, Saudi Arabia
peter.wonka@kaust.edu.sa

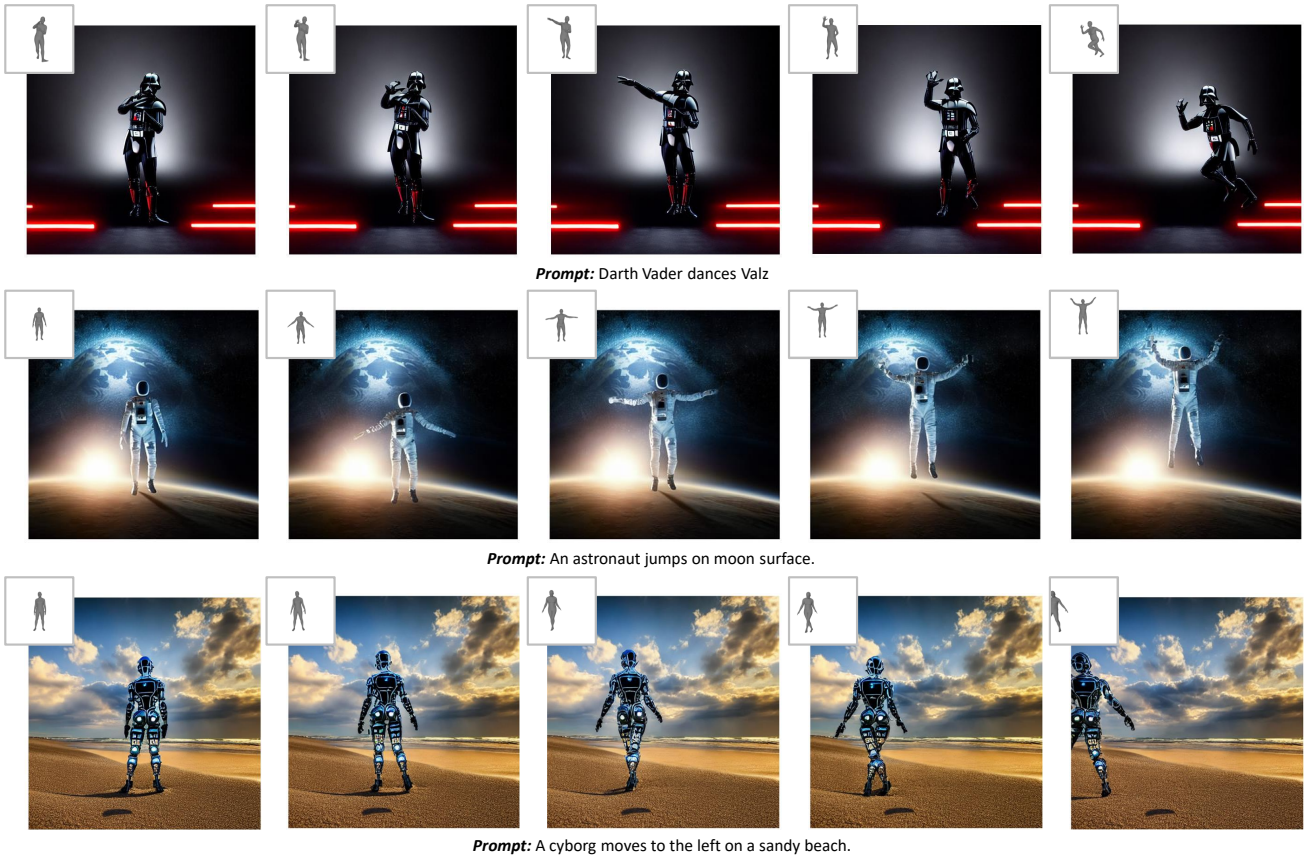


Figure 1. LATENTMAN produces temporally consistent videos of animated characters using pre-trained Motion and Text-to-Image (T2I) diffusion models given only a textual prompt.

Abstract

We propose a zero-shot approach for generating consistent videos of animated characters based on Text-to-Image (T2I) diffusion models. Existing Text-to-Video (T2V) methods are expensive to train and require large-scale video datasets to produce diverse characters and motions. At the same time, their zero-shot alternatives fail to produce temporally consistent videos with continuous motion. We strive to bridge this gap, and we introduce LATENTMAN *that leverages existing text-based motion diffusion models to generate diverse continuous motions to guide the T2I model. To boost the temporal consistency, we introduce the Spatial Latent Alignment module that exploits cross-frame dense correspondences that we compute to align the latents of the video frames. Furthermore, we propose Pixel-Wise Guidance to steer the diffusion process in a direction that minimizes visual discrepancies between frames. Our*

proposed approach outperforms existing zero-shot T2V approaches in generating videos of animated characters in terms of pixel-wise consistency and user preference.

1. Introduction

Generating visual assets of human characters is a prominent task in the realm of image and video synthesis, with many applications in movie production, art, and fashion. This task aims to generate high-quality and diverse images/videos of human characters that adhere to some given conditions, *e.g.* textual prompts and human poses. Text-to-Image (T2I) diffusion models [24–26] revolutionized this endeavor as they can generate high-quality images of human characters conditioned on user-provided textual prompts. ControlNet [36] allowed further control over the generated images through various conditioning signals such as depth maps, human poses, and edge maps.

For generating videos of human characters, Text-to-Video (T2V) diffusion models [2, 3, 11, 28] are evolving rapidly, but there are several complexities associated with them. For instance, learning the motion dynamics (*e.g.* the human body), finding sufficiently large datasets, and fulfilling their excessive computational needs. As an example, the largest publicly available video dataset encompasses only 10 million videos [1], and it requires up to 48 A100-80GB GPUs to train VideoLDM [3] on this dataset. Therefore, a growing direction of research attempts to democratize this task by leveraging T2I models to generate videos in *few-* to *zero-shot* manner.

One category of approaches [8, 23, 32, 34] adopts a Video-to-Video (V2V) scheme that relies on a reference video to generate a target video with modified contents. However, these approaches require the user to provide the reference video, which can be difficult and inconvenient to find. Alternatively, Text2Video-Zero [32] proposed to generate videos based only on a textual prompt where the motion dynamic is simulated by applying translation vectors to the latent codes of the first frame. The temporal consistency was achieved by converting the *self-attention* modules of the T2I UNet, which encodes the visual style, to *cross-frame* attention. This enforces the T2I model to generate video frames that are visually consistent. Nonetheless, the generated videos lack any motion continuity and only show random variations of the same object. Moreover, a closer look at the generated frames shows that the temporal consistency is rather global, and fine details tend to change.

To illustrate this observation, we conduct a controlled experiment where we render a SMPL human model [18] to obtain a depth map of a human. We use this depth map and a textual prompt as conditions for ControlNet to generate a *reference image*. Then, we shift the depth map upwards by 10 pixels to simulate a moving human in a video. We replace self-attention modules with cross-frame attention as in Text2Video-Zero to enforce the T2I model to generate frames with the same style as the reference frame. Figure 2 shows that cross-frame attention successfully preserves the overall style of the frames. However, the fine details of the robot (shown in the insets) tend to change between frames. We find that this is caused by the distributional shift in the latent codes that are responsible for generating the character in the scene as shown on the right of Figure 2.

To this end, we propose a zero-shot approach for generating consistent videos of animated characters based on T2I diffusion models. To produce continuous motion dynamics, we employ text-based human motion diffusion models [31] to generate a sequence of SMPL models given a text prompt. We render these SMPL models to generate a sequence of depth maps that can be used as conditional inputs for ControlNet. This allows generating videos with realistic and continuous animations, unlike Text2Video-Zero. To boost temporal consistency, we compute cross-frame dense correspondence based on DensePose [9], and we use it to align the latent codes between video frames through our *Spatial Latent Alignment* module. We also propose an additional *Pixel-Wise Guidance* strategy that steers the diffusion process in a direction that minimizes the visual discrepancies between frames.

To evaluate the temporal consistency of the generated videos, we introduce the *Human Mean Squared Error* metric that measures the pixel-wise difference of the animated character between consecutive frames. Our proposed approach outperforms Text2Video-Zero on this metric by $\sim 10\%$ and was preferred by 76% of the users in a user study that we conducted.

Our contributions can be summarized as follows:

- We introduce a zero-shot approach for generating videos of animated characters.
- We employ Motion Diffusion Models to generate continuous motion guidance based solely on text.
- We propose the *Spatial Latent Alignment* and *Pixel-Wise Guidance* modules that boost temporal consistency.
- Our approach outperforms existing zero-shot approaches in terms of the *Human Mean Squared Error* metric that we introduce and in terms of user preference.

2. Related Work

We give a brief overview of existing approaches for human video synthesis, Text-to-Video (T2V) diffusion models, and human motion synthesis.

Human Video Synthesis Existing approaches for human video generation are generally limited to specific domains and datasets. For instance, several T2V approaches [19, 28, 35] train on the UCF-101 dataset [30] that includes videos of humans performing 101 diverse actions. However, the generated videos based on this dataset are low res-

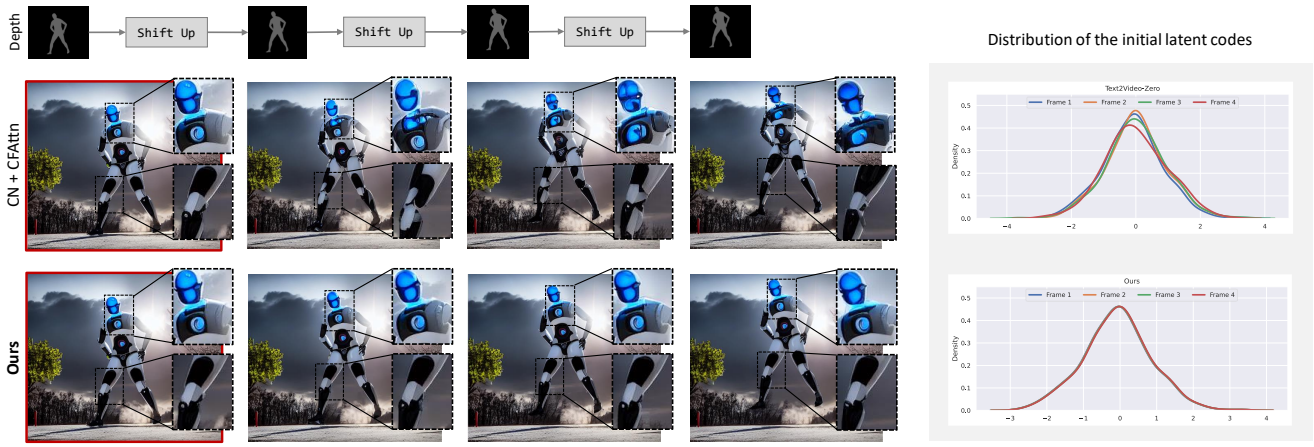


Figure 2. *Cross-Frame Attention* (CFAttn) is adopted by multiple zero-shot T2V approaches to generate globally consistent video frames. However, when the conditioning signal (the depth map) changes, *e.g*. shifted up, the fine details (shown in the insets) tend to vary between frames. We find that this is caused by the distributional shift of the initial latent codes that are aligned with the character, as shown on the plot to the right. Our proposed approach attempts to align the latent codes in a zero-shot manner, eliminating the distribution shift and producing consistent images. *CN refers to ControlNet

olution and lack visual diversity. Another category of approaches focused on generating videos of fashion performers and is trained on fashion datasets [13, 17]. For instance, Text2Performer [13] proposed a decomposed human representation into pose and appearance in the latent space of a variational autoencoder. This representation is used alongside a diffusion-based motion sampler to generate consistent high-resolution videos of fashion performers. Nevertheless, their approach can only generate videos of performers with standardized motions on a simple background.

Recently, several approaches [12, 20, 33] proposed diffusion models for Image-to-Video (I2V) to animate a human character given a subject image and a sequence of poses that are provided by the user. Contrarily, we address the Text-to-Video (T2V) problem that aims to produce diverse videos of animated characters based solely on a textual prompt. It is worth mentioning that the concurrent work [4] shares similarities with our work as it attempts to generate consistent videos given a sequence of UV maps by aligning the latent codes. But we differ from them in that we only require textual prompts as input and that we follow a different strategy for aligning the latents.

Text-to-Video Diffusion Models Text-to-Image (T2I) diffusion models [24–26] excelled in generating highly realistic and diverse images based on textual prompts by harnessing large-scale image datasets [27]. With the lack of similarly large video datasets to train T2V counterparts, a growing direction of research attempts to exploit existing T2I models to generate videos. VideoLDM [3] proposed to transform a pre-trained Stable Diffusion model [25] into a T2V model by introducing a temporal module and a video upsampler that are trained on video data. Similarly, Make-a-video [28] extended DALLE-2 [24] to a T2V model by temporally aligning the decoder and the upsampler on video data. However, these two approaches require excessive GPU resources and large-scale datasets to train.

Tune-a-Video [32] adopts a one-shot paradigm and fine-tunes a pre-trained T2I model to generate a video given a single video/text pair. Nevertheless, this approach *requires a video as an input* in addition to the text prompt, making it more suitable for video editing or Video-to-Video tasks. Text2Video-Zero [15] introduced a purely text-based zero-shot T2V approach that injects motion dynamics into the latents of a T2I diffusion model. Their approach exploited the fact that the output of diffusion models varies under any changes to the latent codes to generate variations of the first frame. However, the generated frames from their approach lack any motion continuity or temporal consistency. In contrast, our approach employs Motion Diffusion Models [6, 31] to generate continuous motion guidance and introduces two strategies for boosting temporal consistency, especially at fine details.

Human Motion Synthesis This task aims to produce animated skeletons (standardized poses) of humans conditioned on textual prompts. Several approaches for human motion synthesis were proposed that benefited from the large datasets for human motions, such as the HumanML3D dataset [10] with approximately 15k diverse motions. T2M [10] proposed a two-stage approach that learns a mapping function between text prompt and motion length. Afterward, a temporal variational autoencoder generates the motion given the predicted length. MDM [31] employed a diffusion model to learn a conditional mapping between text and motion sequences. MLD [6] learns a compact latent

representation to train diffusion models in a more efficient manner. GMD [14] incorporated spatial constraints into the motion diffusion process to add more control over the generated motions. We employ any of these approaches to generate diverse and continuous motion signals to guide a pre-trained T2I diffusion model.

3. Method

In this section, we first describe the existing pipeline for zero-shot Text-to-Video (T2V) diffusion models that is adopted by Text2Video-Zero [15] and MasaCtrl [5]. Then, we explain our proposed approach for generating temporally consistent videos of animated characters.

3.1. Zero-Shot T2V Diffusion Models

The objective of the T2V task is to generate a sequence of N video frames $\mathcal{I} := \{I_1, I_2, \dots, I_N\}$, given a text prompt $\mathcal{T}$. In the zero-shot setting, a pre-trained Text-to-Image (T2I) diffusion model such as Stable Diffusion (SD) [25] is used to generate each frame individually. For better control over the contents of the generated frames, additional conditioning signals $\mathcal{G} := \{G_1, G_2, \dots, G_N\}$, such as human poses, canny edges, and depth maps are incorporated through ControlNet [36] or T2I-Adapters [21].

During inference, the diffusion process is carried out using a denoising model such as DDIM [29], where for each frame i and denoising step t, we compute the previous latent code x^i_{t-1} as well as, a noise-free sample prediction $\hat{x}^{i,t}_0$:

$$\hat{x}^{i,t}_0 = \frac{x^i_t - \sqrt{1-\alpha_t}\,\epsilon^t_\theta(x^i_t, \mathcal{T}, G_i)}{\sqrt{\alpha_t}} \quad , \tag{1}$$

$$x^i_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{x}^{i,t}_0 + \sqrt{1-\alpha_{t-1}-\sigma^2_t}\,\epsilon^t_\theta(x^i_t, \mathcal{T}, G_i) + \sigma_t \epsilon_t \quad , \tag{2}$$

where α_t, σ_t are pre-defined scheduling parameters, ϵ^t_θ is a noise prediction from a trained UNet, and ϵ_t is random Gaussian noise. This process is computed for $T \geq t \geq 0$, and the final image is reconstructed at $t = 0$ using the decoder of a variational autoencoder as $I_i = \mathcal{D}(\hat{x}^{i,0}_0)$. To promote visual consistency, the initial latent code x_T is shared among all frames, and the self-attention modules are replaced with *cross-frame attention*. We refer the reader to [5, 15] for details on cross-frame attention. We use this aforementioned pipeline as a baseline for our approach.

3.2. Zero-Shot Text-to-Animated-Characters

To generate videos of animated characters using T2I models, we need conditioning signals $\mathcal{G}$ to control the generated content. Existing methods [15, 32] extract these signals from a user-provided video. For example, a depth or human pose detector is applied to a video to extract depth maps or human poses. However, this approach has limited control over the generated content and adds the burden of finding a suitable video.

Instead, we propose to employ text-based motion diffusion models [6, 31] to produce a sequence of length N of human skeletons, given the text prompt $\mathcal{T}$. Afterward, we fit a customizable human body model such as SMPL [18] to each of these skeletons, and we render N depth maps from these models to produce conditioning signals $\mathcal{G}^{depth}$. We also compute DensePose [9] for each frame to obtain $\mathcal{P} := \{P_1, P_2, \dots, P_N\}$. This approach eliminates the need for providing a reference video as in [5, 15, 32] and makes the process purely text-driven. An overview of our proposed approach is illustrated in Figure 3.

3.3. Cross-Frame Dense Correspondences

Since we obtained per-frame conditioning signals $\mathcal{G}^{depth}$, we can directly generate the video frames. However, as demonstrated in Figure 2, the output of SD varies under any changes to the conditioning signal, causing the frames to be temporally inconsistent. To alleviate this problem, the latent codes during inference must be spatially aligned, *i.e.*, each body part of the generated character needs to have the same latent code in all frames. To achieve this, we need to compute pixel-wise dense correspondences between frames and use them to propagate the latent codes across frames.

Ideally, the UV maps for the SMPL models or DensePose can be used for this purpose. However, since they need to be downsampled to the resolution of the latent code, the correspondences are lost, and they need to be recomputed. To tackle this issue, we set up a dense correspondence problem between each two consecutive frames based on the DensePose embeddings. We opt for DensePose rather than the UV maps as the former divides the human body into parts, making the correspondence problem cheaper to solve. We denote the DensePose embedding for frame i as $P_i = [L_i, U_i, V_i]$, where $P_i \in \mathcal{P}$, L_i has pixel-wise labels for body parts in the range of $[0, 24]$, and U_i, V_i are UV-coordinates in the range of $[0, 255]$. For each body part j, we define a set of pixels that belong to that part as $Q^j_i := \{q \mid L_i(q) = j\}$. We form a feature vector for each $q \in Q^j_i$ and we arrange them in the rows of matrix $\hat{P}^j_i$:

$$\hat{P}^j_i[q] = \begin{bmatrix} U^j_i(q) & V^j_i(q) & E^j_i(q) \end{bmatrix} \quad , \tag{3}$$

where $E^j_i[q]$ is the euclidean distance between pixel q and the centroid of body part j. This term encourages the matching of pixels that are spatially close.
For each two consecutive frames i and $i-1$, we compute a cost matrix C between $\hat{P}^j_i$ and $\hat{P}^j_{i-1}$ as:

$$C[q, s] = \| \hat{P}^j_i[q] - \hat{P}^j_{i-1}[s] \|_2 \quad , \tag{4}$$

where $q \in Q^j_i, s \in S^j_{i-1}$, and $S^j_{i-1} := \{s \mid L_{i-1}(s) = j\}$. Then, we find the correspondences by solving a linear assignment problem over C using the Hungarian algorithm

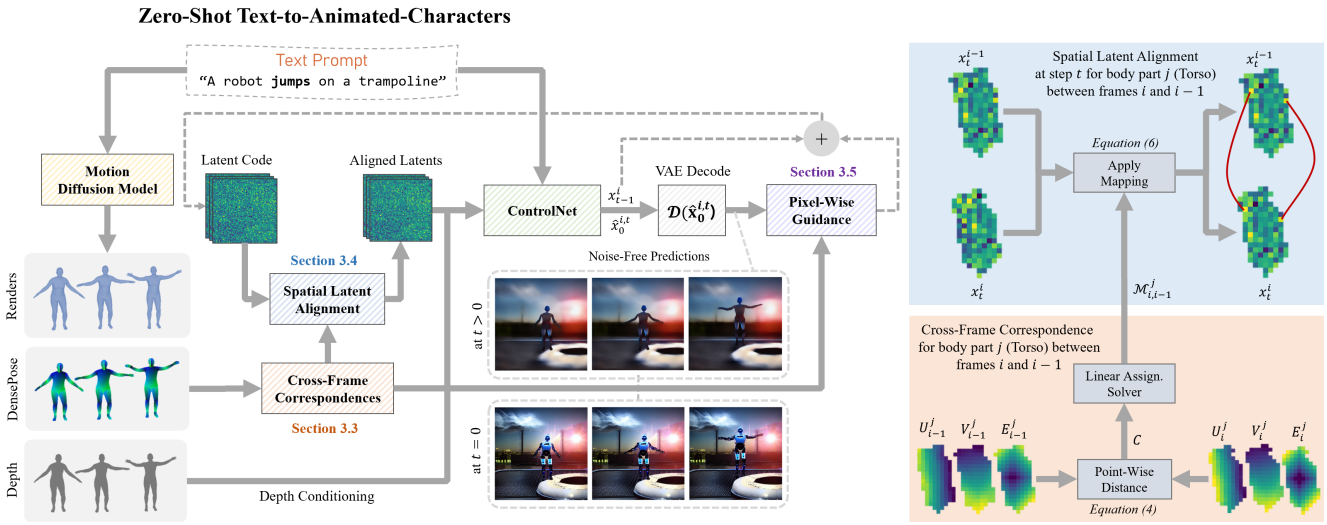


Figure 3. An overview of our proposed approach. Given a text prompt $\mathcal{T}$, a motion diffusion model [31] produces a sequence of human skeletons that we use to obtain frame-wise depth maps and DensePose [9]. The former is used as guidance for ControlNet [36], while the latter is used to compute cross-frame correspondences. These correspondences are employed by the Spatial Latent Alignment and the Pixel-Wise Guidance modules to boost temporal consistency. The orange block shows an illustration of how we compute cross-frame correspondences between two frames for the "torso" body part based on DensePose. The blue block shows how we employ these correspondences to spatially align the latents to promote consistent synthesis.

[16], which assigns each pixel to the closest match based on the UV coordinates and the spatial location. This produces an *injective* mapping for each body part j between frames i and $i-1$ as:

$$\mathcal{M}^j_{i,i-1} := \{(q,s) \,\forall\, q \in Q^j_i, s \in S^j_{i-1}\} \ , \tag{5}$$

An illustration for this procedure is shown in Figure 3. Finally, All body parts are then combined into a total body mapping as $\mathcal{M}_{i,i-1} = \cup_j \mathcal{M}^j_{i,i-1}$.

3.4. Spatial Latent Alignment

To achieve temporal consistency, we aim to align the latents between the video frames. We compute correspondence mappings from Section 3.3 between each two consecutive frames based on DensePose embeddings $\mathcal{P}$ that are downsampled to 64×64 to match the resolution of the latent codes. For frames i and $i-1$, the latent code x^i_t is updated with values from x^{i-1}_t based on the computed mapping as:

$$x^i_t[q] = x^{i-1}_t[s] \quad \forall\,(q,s) \in \mathcal{M}_{i,i-1} \ , \tag{6}$$

This operation will copy some parts of the latent code from frame $i-1$ to the correct spatial location in frame i, promoting temporal consistency. Note that we only apply this operation at the first 40% of the diffusion steps that encompass the generation of the main structures of the scene.

3.5. Pixel-Wise Guidance

The resolution of the latents in SD is $1/8$ of that of the generated images. Consequently, even after spatially aligning the latents in the Section 3.4, some high-resolution details will vary between the video frames. To alleviate this problem, we propose a Pixel-Wise Guidance strategy inspired by classifier guidance in diffusion models [22]. First, we compute a mapping $\mathcal{M}_{i,i-1}$ from Section 3.3 between each two consecutive frames i and $i-1$. At a given diffusion step t, we reconstruct the RGB predictions using the VAE

Algorithm 1 Zero-Shot Animated Characters Synthesis

Require: $N, T \in \mathbb{N}$, $\delta \in \mathbb{R}$, text prompt $\mathcal{T}$, $A := [a_1, a_2]$, $B := [b_1, b_2]$ ControlNet (`CN`), DDIM (`DDIM`), Motion Diffusion Model (`MDM`), Spatial Latent Alignment (`ALIGN`), Pixel-Wise Refinement (`REFINE`)
Output: $\mathcal{I} := \{I_1, I_2, \ldots, I_N\}$

```
G^depth, P ← MDM(T)                          ▷ Depth maps, DensePose
x_T ~ N(0, I)
for i = 1, ..., N do:
    for t = T, T − 1, ..., 0 do:
        if i > 1 and t ∈ A then              ▷ Spatial Latent Alignment
            x^i_t ← ALIGN(x^i_t, P[i], P[i − 1])
        end if
        ε^t_θ ← CN(x^i_t, T, G^depth[i], t)
        x^i_{t−1}, x̂^{i,t}_0 ← DDIM(x^i_t, ε^t_θ)
        if i > 1 and t ∈ B then              ▷ Pixel-Wise Guidance
            ω_i ← REFINE(x̂^{i,t}_0, P[i], P[i − 1])
            x^i_{t−1} ← x^i_{t−1} − δ ∇_{x^i_t} ω_i
        end if
    end for
end for
```

Figure 4. A qualitative comparison between our proposed approach and the baseline Text2Video-Zero [15]. Our approach is able to generate consistent shapes and textures compared to the baseline. The reference frame is the first frame of the video that defines the appearance of the character.

decoder as $X_t^i = \mathcal{D}(\hat{x}_0^{i,t})$ and we compute the L2 difference between all pixel pairs in $\mathcal{M}^{i,i-1}$:

$$\omega_i = \sum_{q,s} (X_t^i[q] - X_t^{i-1}[s])^2 \quad \forall\, (q,s) \in \mathcal{M}_{i,i-1} \ , \quad (7)$$

Finally, we compute the gradient of ω_i with respect to x_t^i, and we use it to update x_{t-1}^i:

$$x_{t-1}^i = x_{t-1}^i - \delta \, \nabla_{x_t^i} \, \omega_i \ . \quad (8)$$

where δ is a scaling factor. This steers the diffusion process in the direction that minimizes w_i.

Note that we apply this process on the resolution of 256×256 rather than the full resolution of 512×512 as the latter would be computationally expensive using the Hungarian algorithm with cubic complexity.

4. Experiments

We evaluate our approach based on two baselines that adopt cross-frame attention. The first baseline is MasaCtrl [5], which is an image editing method that can be used to generate a sequence of consistent images, and the second is Text2Video-Zero [15], a zero-shot approach for video synthesis. Note that Text2Video-Zero has two variations: a

Figure 5. Impact of motion guidance in our approach compared to the motion dynamics in Text2Video-Zero [15]. Motion guidance produces videos that adhere to the prompt in contrast to Text2Video-Zero, which produces random variations of the scene.

condition-free version and a conditional one based on ControlNet. We compare mainly against the latter, but we provide some examples for the former as well.

4.1. Implementation Details

For both baselines, we use a pre-trained Stable Diffusion [25] version 1.5 with ControlNet [36] depth control to generate 512×512 video frames. For inference, we employ the DDIM sampler [29] with a linear schedule. We use $T = 100$ inference steps for Text2Video-Zero and $T = 50$ for MasaCtrl. We empirically choose the guidance factor $\delta = 0.01$ in Equation (8), $A = [0, 39], B = [20, 69]$ for Text2Video-Zero, and $A = [0, 19], B = [20, 39]$ for MasaCtrl in Algorithm 1. For motion synthesis, we use the official implementation of *Motion Diffusion Model (MDM)* [31] with some modifications for rendering and computing DensePose [9]. We conduct all experiments on a single NVIDIA A100 GPU except for Gen-2 [7], where we use the official demo. Our code will be made publicly available.

4.2. Qualitative Results

Zero-Shot Comparison Figure 4 shows a qualitative comparison between depth-conditioned Text2Video-Zero and our proposed approach. The figure shows that cross-frame attention adopted by Text2Video-Zero is not sufficient to preserve the fine details of the generated characters. As the conditional depth map changes, the object gets distorted (*e.g.* the robot torso and legs in the first row), or the texture changes (*e.g.* the pants in the second row become shorts). On the other hand, our approach successfully maintains the fine details of the generated characters across all frames.

Motion Guidance Significance To demonstrate the impact of motion guidance produced by MDM, we compare our approach against Text2Video-Zero with no depth conditioning, which produces video by injecting motion dynamics into the latent codes. Figure 5 shows that Text2Video-Zero produces random variations of the scene that do not adhere to the motion in the prompt. For example, the astronaut in the top row is just floating and not jumping, and Ironman in the second row does not move but changes pose. On the other hand, our approach produces consistent videos that adhere to the motion in the prompt.

	$H_{MSE} \downarrow$	User Preference [%]
MasaCtrl [5] [ICCV23]	88.19	34 %
Ours	**79.88**	**66 %**
Text2Video-Zero [15] [ICCV23]	84.87	24 %
Ours	**76.41**	**76 %**

Table 1. Quantitative comparison between our proposed approach and two baselines. The error reduction percentage is shown

Trained T2V Comparison We also provide a comparison against the trained T2V model, Gen-2 [7], in Figure 6. In the first column, Gen-2 fails to produce a video of a robot jumping on a trampoline, and the robot morphs into a sphere. Our approach manages to produce a video for this uncommon scenario as the generation of the motion and the style are decoupled. In the second column, Gen-2 produces a good video of a skier with rich video dynamics. However, the skier loses his backpack after a few frames and deforms by the end of the video. Our approach produces a consistent video but with less background dynamics.

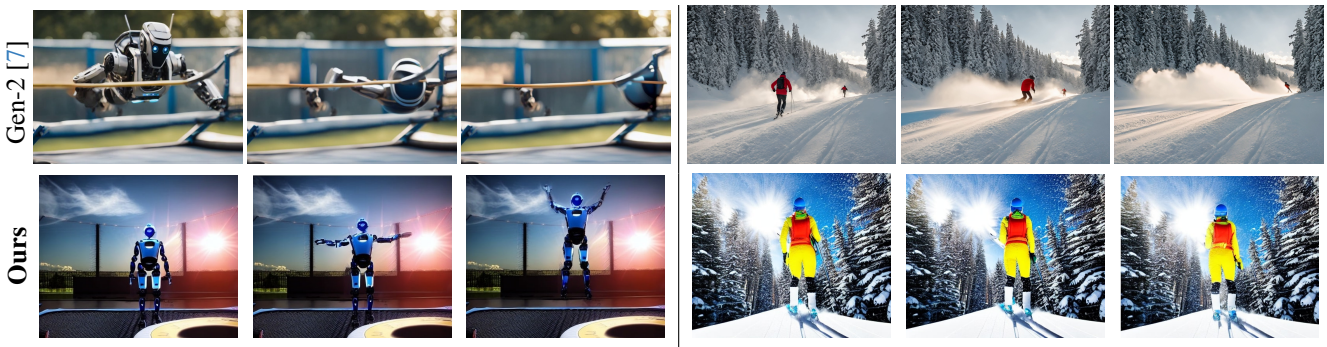
4.3. Quantitative Results

To numerically evaluate the generated videos, we introduce a new metric for temporal consistency and perform a user study. We denote the new metric as the *Human Mean Squared Error* H_{MSE}, and it compares the pixel-wise values of the generated characters in every two consecutive frames. We employ the computed cross-frame dense correspondences $\mathcal{M}$ from Section 3.3, and we compute the mean squared error (MSE) between corresponding pixels:

$$H_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\mathcal{M}_{i,i-1}|} \sum_{q,s} (I_i[q] - I_{i-1}[s])^2 \quad \forall\, (q,s) \in \mathcal{M}_{i,i-1} \tag{9}$$

where I_i, I_{i-1} are the final generated frames in $\mathcal{I}$.

We generated 10 videos of diverse motions and characters and we computed the proposed metric and performed the user study on them. Table 1 shows that our approach outperforms both baselines in terms of H_{MSE} by $\sim 9-10\%$, which demonstrates that the generated characters are temporally more consistent. For the user study, users are asked to select between two videos; one is produced by the



Prompt: "A robot jumps on a trampoline"

Prompt: "A skier running on a snowy road"

Figure 6. A comparison against the trained T2V model Gen-2 [7].

	$H_{MSE} \downarrow$	Runtime (s)
Baseline	84.87	28
with SLA	78.90 (-7.0 %)	30
with PWG	82.73 (-2.5 %)	49
with LSA + PWG	76.41 (-10.0 %)	50

Table 2. An ablation study for different components of our proposed approach. *SLA*: Spatial Latent Alignment in Section 3.4, *PWG*: Pixel-Wise Guidance in Section 3.5. Runtime is reported for generating a 8-frames video.

baseline and the other by our approach. Table 1 shows that 76% and 66% of the users (based on 23 users) preferred the videos generated by our approach over Text2Video-Zero and MasaCtrl baselines, respectively.

4.4. Ablation Study

We provide an ablation study in Table 2 to show the contribution of each component in our proposed pipeline to the overall performance. The Spatial Latent Alignment module contributes the most to the overall improvement and improves by 7.0% over the baseline. This indicates that aligning the latents plays a crucial role in achieving temporal consistency. Pixel-Wise Guidance in Section 3.5 improves over the baseline by 2.6% as it is mainly focused on the fine details. The two components combined achieve a joint improvement of 10 % compared to the baseline.

4.5. Limitations and Failure Cases

Since we employ ControlNet with depth conditioning for generating the video frames, our approach is also bounded by its limitations. For example, the top row of Figure 7 shows an example where ControlNet fails to produce a realistic left arm and leg when they intersect in the depth map. Another source of failure is mismatches when computing the correspondence mapping in Section 3.3, which can lead to some artifacts. It is also worth mentioning that Pixel-Wise Guidance imposes high GPU memory usage due to computing gradients with respect to the latent codes. However, employing the Spatial Latent Alignment solely still achieves remarkable improvement over the baseline as shown in Table 2 with no GPU memory overhead.

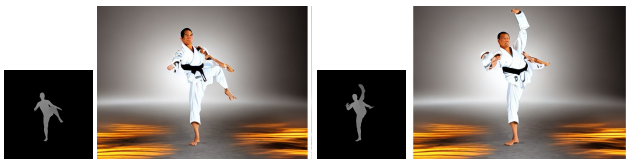


Figure 7. Examples of a failure case

5. Conclusion and Future Work

We introduced a new paradigm for generating consistent videos of animated characters in a zero-shot manner. We employed text-based motion diffusion models to provide continuous motion guidance that we utilized to generate video frames through a pre-trained T2I diffusion model. This allowed generating videos of diverse characters and motions that existing T2V methods struggled to produce. We also demonstrated that our approach produces temporally consistent videos achieved through the proposed Spatial Latent Alignment and Pixel-Wide Guidance modules. These two modules can benefit other approaches that adopt cross-frame attention and latent diffusion models in general. For future work, the cross-frame dense correspondences can be improved for better latent alignment. Furthermore, video dynamics can be incorporated into the background for enhanced realism.

References

[1] 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. 2

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[4] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. *arXiv preprint arXiv:2312.01409*, 2023. 3

[5] 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. *arXiv preprint arXiv:2304.08465*, 2023. 4, 6, 7

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3, 4

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 7, 8

[8] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 2

[9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2, 4, 5, 7

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 3

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[12] 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. 3

[13] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[14] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Gmd: Controllable human motion synthesis via guided diffusion models. *arXiv preprint arXiv:2305.12577*, 2023. 4

[15] 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. *arXiv preprint arXiv:2303.13439*, 2023. 3, 4, 6, 7

[16] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[17] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 4

[19] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023. 2

[20] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 3

[21] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 4

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 5

[23] 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 (ICCV)*, pages 15932–15942, 2023. 2

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 3

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 4, 7

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep

language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[28] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 4, 7

[30] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[31] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 4, 5, 7

[32] 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. 2, 3, 4

[33] 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. In *arXiv*, 2023. 3

[34] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *arXiv preprint arXiv:2306.07954*, 2023. 2

[35] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 2

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 4, 5, 7