Towards motion from video diffusion models

Paul Janson[1,2], Tiberiu Popa[1], and Eugene Belilovsky[1,2]

[1] Concordia University, Montreal , Canada
[2] MILA Quebec AI Institute, Montreal, Canada
p_janso@live.concordia.ca

Abstract. Text-conditioned video diffusion models have emerged as a powerful tool in the realm of video generation and editing. But their ability to capture the nuances of human movement remains under-explored. Indeed the ability of these models to faithfully model an array of text prompts can lead to a wide host of applications in human and character animation. In this work, we take initial steps to investigate whether these models can effectively guide the synthesis of realistic human body animations. Specifically we propose to synthesize human motion by deforming an SMPL-X body representation guided by Score distillation sampling (SDS) calculated using a video diffusion model. By analyzing the fidelity of the resulting animations, we gain insights into the extent to which we can obtain motion using publicly available text-to-video diffusion models using SDS. Our findings shed light on the potential and limitations of these models for generating diverse and plausible human motions, paving the way for further research in this exciting area.

Keywords: Diffusion models · Human motion generation · Digital Humans

1 Introduction

Video generative models [8, 14, 30, 48, 57, 59] have been shown and claimed to be potential tools for simulating the world. Recent advancements, such as those demonstrated by [30], highlight the capabilities of video diffusion models trained on vast datasets to generate realistic and diverse visual content. These impressive capabilities inspired us to question the potential of open-source counterparts for the specific task of generating human motion animations from natural language input. We sought to determine if the current state-of-the-art open-source models can be used with score distillation to generate human motion effectively.

Human motion generation from textual instructions is a well-studied domain [18,33,54,55], with recent methods often involving diffusion models trained on motion capture (MoCap) data [46, 55]. However, the scarcity of MoCap data compared to the abundance of video data presents a scalability challenge. Extracting animation directly from videos, mirroring the success seen in 3D asset generation [11, 24, 29, 60], represents an ambitious yet necessary goal.

We define our problem as determining the correct sequence of joint rotations necessary to produce realistic human motion described by a prompt. To

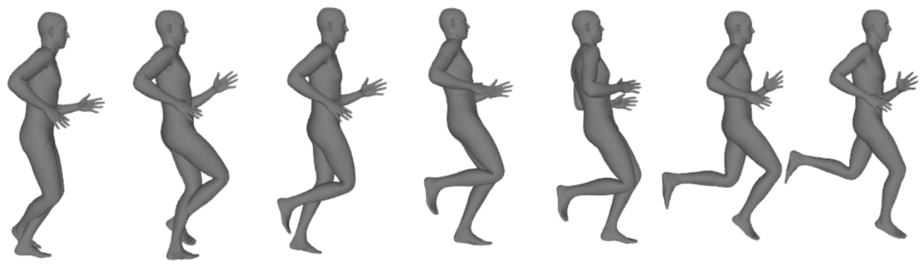

Fig. 1: Human motion sequence resembling running generated using text-to-video model. The figure illustrates that the current video models can generate realistic motion for commonly occuring human activity such as running

achieve this, we utilize the widely adopted SMPL-X [32] digital human template model to render a character. This character is animated through an optimization process that iteratively updates a multi-layer perceptron (MLP) to predict the corresponding pose parameters. We guide this optimization process using a video diffusion model, which provides feedback on the realism of the generated motion.

As shown in Figure 1, these models demonstrate a proficiency in generating animations for common actions such as running. However, when tasked with depicting uncommon or rare human movements, their performance under score distillation reveals limitations. While they can produce visually appealing results for familiar actions, their ability to capture the nuances of less frequent movements remains a challenge. Our main contributions are as follows:

- We propose a differentiable video generation pipeline: **MotionDistill** that leverages text-to-video diffusion models to generate human motion.
- We conduct an evaluation of ModelScope [59], ZeroScope [6], and VideoCrafter [7, 8]'s capabilities in generating both common and uncommon human motion.
- We ablate our analysis to the latent space of these models to test the effectiveness of SDS

2 Related Works

Diffusion models in content creation: Recent advances in text-to-image [40, 43] foundation models [4] have catalyzed exploration into their application for three-dimensional (3D) asset synthesis [50] and editing [62]. Early works, particularly with CLIP [38]-based models , leveraged joint image-text embeddings to generate 3D assets directly from text prompts [17, 18, 29, 44, 58, 61]. However, the emergence of diffusion models [9, 51] and the introduction of SDS [37] marked an important shift. SDS enabled the extraction of 3D assets [11, 24, 37, 53, 60] as NeRFs [28], Meshes [5, 23, 47] and Gaussian splats [19]. [49] integrated a temporal dimension, facilitating the generation of animated 2D [10] and 3D assets from video diffusion models. [3, 25, 63], introduced hybrid SDS methodologies and alternative representations like Gaussian splattings [25] to enhance the fidelity and motion quality of generated assets. However, these approaches focused on

open-ended generation and did not specifically address the potential of video models to generate diverse human motions.
Text to Human motion generation: The generation of human motion guided by textual descriptions is a well-established area of research. Initial approaches explored to model it as machine translation [1, 36] and joint cross-modal mappings [2,12] to address this challenge. Subsequent works leveraged motion capture datasets [27,35] to train models capable of generating human motion as sequences of poses. Variational Autoencoders (VAEs) [21] were employed in [13, 34]. The concept of a shared latent space with CLIP [38] was introduced in [54]. Human motion diffusion model [55] pioneered the application of diffusion-based modeling to human motion generation, enabling ancestral sampling in the motion space, which subsequently led to the utilization of SDS [37] in [46] for generating extended motion sequences. These prior methods are limited by motion capture data, operating solely in the rotation space. Our approach ventures into the pixel space models, exploring their potential in motion generation.

3 Background

Human template models - SMPLx: The Skinned Multi-person Linear(SMPL) [26, 32, 41, 45] family of models comprises articulated human body models parameterized by the shape and pose parameters. Among these, We use SMPLx because of its comprehensive representation of the human body. It is defined by the function $M(\theta, \beta, \phi) : \mathbb{R}^{|\theta|\times|\beta|\times|\psi|} \rightarrow \mathbb{R}^{3N}$ [32] given by Equation 1.

$$M(\beta, \theta, \psi) = LBS(T_p(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}) \tag{1}$$

This function takes body parameters as input, performs linear blend skinning, and outputs a mesh with $N = 10,475$ vertices. Here $\theta \in \mathbb{R}^{3(K+1)}$ denotes the body pose parameters, $\beta \in \mathbb{R}^{|\beta|}$ denotes the shape parameters and $\psi \in \mathbb{R}^{|\psi|}$ denotes the facial expression parameters. $\mathcal{W}$ is the skinning weights and J is the joint regressor. The pose parameters θ can be further divided into θ_b (body joints pose), θ_f (jaw pose) and θ_h (finger pose). While SMPLx accounts for $K = 54$ body joints, our approach optimizes only the major body joints $K_b = 21$ focusing exclusively on θ_b to constraint our scope. The template body $T_p(\beta, \theta, \psi)$ is given by Equation 2. Here $\hat{T}$ is the template mesh, B_S, B_E, B_P denotes the blend shape functions corresponding to shape, expression, and pose.

$$T_p(\beta, \theta, \psi) = \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\theta; \mathcal{P}) \tag{2}$$

Score Distillation Sampling: Score Distillation Sampling (SDS) [37] is a method employed to leverage large-scale diffusion models [15, 51] for training compact parametric image generators. It utilizes the score function [52] of the diffusion model to derive gradient directions for updating the generator, iteratively aligning it with provided textual prompts. Our formulation of SDS slightly deviates from the original by employing a latent diffusion model [39].Given a pre-trained latent diffusion model ϕ with its denoising UNet $\hat{\epsilon}_\phi(z_t; y, t)$ [42], text prompt y and an image generator parameterized by θ. The gradients needed to

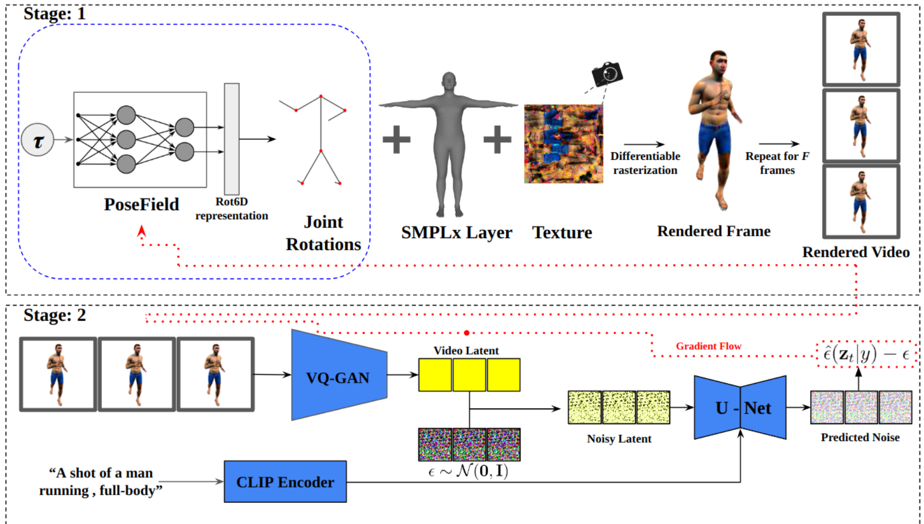

Fig. 2: Our study consists of two stages. Stage:1 (top) Joint rotations required to animate the character are generated using PoseField. Passed through SMPLx Layer to get the final mesh which is then rasterized using a differentiable renderer. We use a random camera and a predetermined texture. This is repeated for F frames to obtain the video. Stage:2 (bottom) Rendered video is encoded to the latent space of the diffusion model then random noise is added to the latent. Unet of the Video model is used to predict the added noise and the gradients are estimated by SDS.

update the generator are obtained by Equation 3. Here t is the randomly sampled diffusion timestep, and ϵ is the randomly generated noise. z is the generated image encoded to the latent space of the diffusion model and z_t is the noised image and $w(t)$ is a weighting function. [39, 51].

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, z) = E_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(z_t; y, t) - \epsilon)\frac{\partial z}{\partial \theta}] \tag{3}$$

4 Method

In this section, we detail our proposed methodology. The central challenge we address is as follows:

Given a text prompt $\mathbf{y}$ describing an action, we aim to determine the optimal joint rotations θ_b for each frame $\{\theta_{b_1}, \theta_{b_2}, ...\}_F$ of an F-framed video that aligns with the given prompt. To tackle this, we model the rotations implicitly using a neural network "PoseField" $P_\alpha : \tau \to \theta_b$ parameterized by α. It is a Multi-Layer Perceptron (MLP) comprising two hidden layers. This network takes the frame id τ as input and outputs the corresponding body pose parameter θ_b. Consequently, our task shifts from finding the optimal rotation set to determining the optimal parameters α^* of the network P_α.

As illustrated in Figure 2, our approach employs a two-stage process.

1. **Video generation:** We utilize a differentiable pipeline to synthesize a video of the animation sequence.
2. **Gradient Estimation:** We leverage video diffusion models to estimate gradients for updating the PoseField parameters.

4.1 Stage 1: Generating multi-view video of the animation sequence

In this stage, we generate the video of the animation sequence frame by frame and subsequently concatenate these frames to form a complete video. We encode the frame identifier τ with positional encoding [56] and infer the PoseField P_α to obtain the SMPL-X body pose parameter $\boldsymbol{\theta}_b = P_\alpha(\tau)$ corresponding to the current frame. Using this, we derive the mesh $M(\boldsymbol{\beta}, [P_\alpha(\tau), \boldsymbol{\theta}_f, \boldsymbol{\theta}_h], \boldsymbol{\psi})$. It is important to note that only α is the trainable parameter here, while all other values are reused across frames and training iterations. We render the character from a randomly selected camera position sampled from a circular trajectory around the mesh. Given the camera trajectory C and the rendering function R, we obtain the projection matrix π and render the current frame as $I_\alpha = R(\pi(M))$. This process is repeated for F frames, generating a series of frames $I_1, I_2, \ldots, I_F$. We concatenate these frames to obtain the video V_α from a certain view that will be used as input to the next stage for gradient estimation.

4.2 Stage 2: Estimating gradients for update

The generated video is then used to estimate the gradient using Score Distillation Sampling (SDS) [37]. We employ the temporal variation of SDS as proposed in [49]. Since we use Latent Diffusion Models, we first encode the video to the latent space of the video diffusion model: $Z_\alpha = E(V_\alpha)$. We then add noise to the video latent according to the noise schedule: $Z_{(\alpha,\sigma,\epsilon)} = \sqrt{1-\sigma^2}Z_\alpha + \sigma\epsilon$, where ϵ is randomly generated noise and $\sigma \in (0, 1)$ is the noise level. Gradients are then computed by Equation 4. Here λ_{SDS_t} is a hyperparameter that controls the effect of SDS.

$$\nabla_\theta \mathcal{L}_{SDS-T} = \lambda_{SDS_t} E_{\sigma,\epsilon}[w(\sigma)\hat{\epsilon}(Z_{\alpha,\sigma,\epsilon}|y,\sigma) - \epsilon)\frac{\partial Z_\alpha}{\partial \alpha}] \tag{4}$$

Regularization: To encourage smoothness in the generated motion, we introduce a regularization constraint. We add the following loss term, which minimizes the difference between the body poses of consecutive frames. Here λ_{reg} is a regularization coefficient.

$$\mathcal{L}_{reg} = \lambda_{reg} \sum_{i=1}^{F-1} (\theta_{b_{i+1}} - \theta_{b_i}) \tag{5}$$

SDS from Stable diffusion model: Following recent works [3,25,49] and considering the fact that video diffusion models often lack visual quality compared to their image counterparts, we additionally estimate gradients using the standard SDS with an Image Diffusion Model. We treat each video as a batch of images and calculate the gradients using Equation 3.

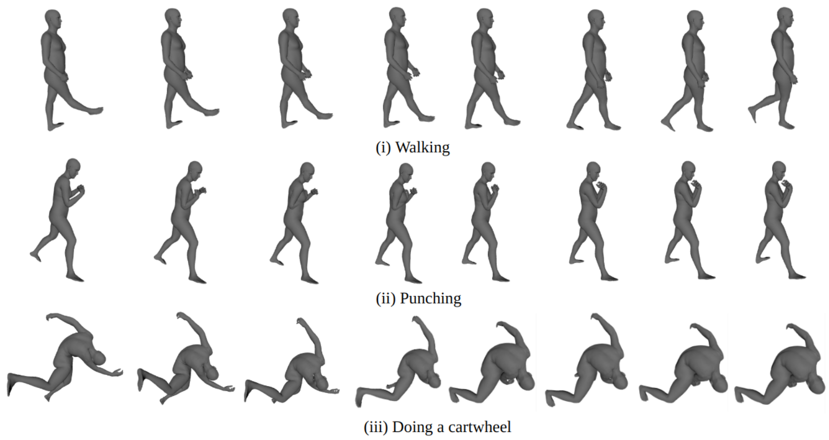

Fig. 3: Our results for different motions: All results are obtained by using the model VideoCrafter [8]. (i) Walking motion is one of the best cases in addition to running (Fig:1). (ii) Punching is a semi-failure case (iii) Cartwheel is an extreme failure case.

5 Experiments and results

Evaluating MotionDistill: Figure 3 showcases the capabilities of our method across various prompts. Our approach excels at generating plausible motion for common human activities, as demonstrated by the "walking" and "running" examples (Figure 3(i) and 1). However, challenges arise with less frequent actions. The "punching" prompt (Figure 3(ii)) resulted in partial motion generation, with limited hand movement. This shows for a semi-failure case. Furthermore, the model struggles with uncommon activities like "doing a cartwheel" (Figure 3(iii)), failing to produce even a reasonable pose giving the extreme failure case [3].

Ablating MotionDistill: It is challenging to determine if the failure cases comes from the SDS, the faithfulness of the video diffusion model, or the way we're representing motion and render it To gain deeper insights into the ability of SDS and our video models to produce faithful motion, we optimize our SDS objective with the video diffusion model directly in their video latent space. We first render the initial video using Stage 1 of our pipeline. Then detach this rendering from the optimization process and directly optimize the latents $\boldsymbol{Z}$. The results are shown in Figure 4.

Our experiments reveal performance discrepancies among three open-source video diffusion models, particularly concerning the generation of common versus rare human motions. This suggests a potential bias towards common activities in the training data. To illustrate this, we used two prompts for all three models: "running" (a common action) and "punching" (less frequent). As shown in Figure 4, the models generated more natural motion sequences for "running." In contrast, the "punching" frames lacked variation, highlighting the limitations

[3] Check supplementary video at https://github.com/Pauljanson002/human-eccv for sample clips

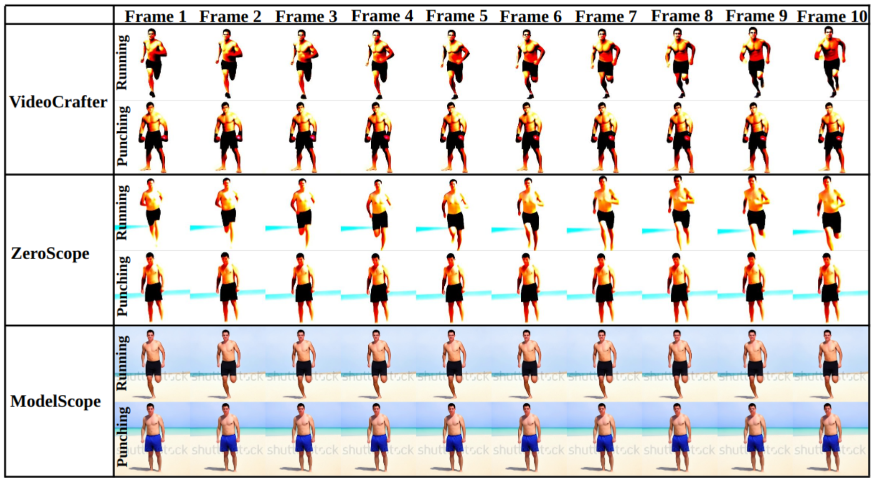

Fig. 4: Visualization of the optimized latent when given the two actions as prompts. Generated videos in the top row of each video model denote the action "running". The bottom row of each denotes the action "punching". Clearly, the top rows of each model show a more natural motion. VideoCrafter [8] demonstrates a higher degree of realism in both actions compared to other models

of all three models with less common motions. We also observed that recent models [8] provide a much more realistic motion than older models [6, 59]. This is evident, if we compare the top rows of each model, ModelScope [59] seems to have very little variation across frames even for "running". Indeed we observe when the model has challenges with the latent generation it also struggles in the case of fitting PoseField, suggesting the issue lies not in the representation of the body motion but in the video diffusion model.

Implementation details: We implemented the pipeline in PyTorch [31] and used nvdiffrast [22] for differentiable rendering. We conducted experiments in NVIDIA A6000 GPU. We used Adam [20] optimizer with a learning rate of 5e-4 for 10,000 iterations. Experiments maintained a CFG [16] scale of 100. We initialized the PoseField to output the mean pose of the target motion and constrained its output within a range defined by three times the standard deviation of the target pose parameters. We set total frames F=10, λ_{reg}=1e-3 , λ_{SDS_t}=1e-3.

6 Conclusion

Video diffusion models can generate human motion from text, but their performance differs a lot between familiar and rare actions. Notably, more recent models like VideoCrafter2 [8] outperform earlier ones [6, 59]. These findings underscore the need for further research to improve the diversity and quality of human motion generation, particularly for less frequent or complex actions. We hypothesize that our method when combined with more powerful text-to-video foundation models can become increasingly more effective. Our work stands as proof of concept in this domain and studies the strengths and limitations of current open-source video diffusion models. Additionally, studying human motion understanding as an emergent behavior of a video diffusion model will be a promising future direction.

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: Generative adversarial synthesis from language to action. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 5915–5920. IEEE (2018) 3
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019) 3
3. Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L., Wonka, P., Tulyakov, S., Park, J.J., Tagliasacchi, A., Lindell, D.B.: 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7996–8006 (June 2024) 2, 5
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N.S., Chen, A.S., Creel, K.A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L.E., Goel, K., Goodman, N.D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T.F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M.S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S.P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J.F., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y.H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K.P., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M.A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the opportunities and risks of foundation models. ArXiv (2021), https://crfm.stanford.edu/assets/report.pdf 2
5. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 958–968 (2024) 2
6. Cerspense: Zeroscope modelcard, https://huggingface.co/cerspense/zeroscope_v2_576w 2, 7
7. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: Videocrafter1: Open diffusion models for high-quality video generation (2023) 2
8. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024) 1, 2, 6, 7
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 2
10. Gal, R., Vinker, Y., Alaluf, Y., Bermano, A., Cohen-Or, D., Shamir, A., Chechik, G.: Breathing life into sketches using text-to-video priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4325–4336 (2024) 2

11. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems **35**, 31841–31854 (2022) 1, 2
12. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1396–1406 (2021) 3
13. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022) 3
14. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 1
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021) 7
17. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.00094, http://dx.doi.org/10.1109/CVPR52688.2022.00094 2
18. Karthikeyan, A., Ren, R., Kant, Y., Gilitschenski, I.: Avatarone: Monocular 3d human animation. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 3635–3645 (2024), https://api.semanticscholar.org/CorpusID:267751541 1, 2
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 2
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Int. Conf. Learn. Represent. (2014) 7
21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3
22. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics (ToG) **39**(6), 1–14 (2020) 7
23. Liao, T., Yi, H., Xiu, Y., Tang, J., Huang, Y., Thies, J., Black, M.J.: Tada! text to animatable digital avatars. In: 2024 International Conference on 3D Vision (3DV). pp. 1508–1519. IEEE (2024) 2
24. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.00037, http://dx.doi.org/10.1109/CVPR52729.2023.00037 1, 2
25. Ling, H., Kim, S.W., Torralba, A., Fidler, S., Kreis, K.: Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8576–8588 (2024) 2, 5
26. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015) 3

27. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.: Amass: Archive of motion capture as surface shapes. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (Oct 2019). https://doi.org/10.1109/iccv.2019.00554, http://dx.doi.org/10.1109/ICCV.2019.00554 3
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020) 2
29. Mohammad Khalid, N., Xie, T., Belilovsky, E., Popa, T.: Clip-mesh: Generating textured meshes from text using pretrained image-text models. In: SIGGRAPH Asia 2022 conference papers. pp. 1–8 (2022) 1, 2
30. OpenAI: Video generation models as world simulators, https://openai.com/index/video-generation-models-as-world-simulators/ 1
31. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Adv. Neural Inform. Process. Syst. (2019) 7
32. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019) 2, 3
33. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10965–10975 (2021), https://api.semanticscholar.org/CorpusID:233210075 1
34. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022) 3
35. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016) 3
36. Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Robotics and Autonomous Systems **109**, 13–26 (2018) 3
37. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023) 2, 3, 5
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 3
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. 2022 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1 (2021) 3, 4
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2
41. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022) 3
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015) 3

43. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022) 2
44. Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.Y., Fumero, M., Malekshan, K.R.: Clip-forge: Towards zero-shot text-to-shape generation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.01805, http://dx.doi.org/10.1109/CVPR52688.2022.01805 2
45. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7763–7772 (2019) 3
46. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: The Twelfth International Conference on Learning Representations (2024) 1, 3
47. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 2
48. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022) 1
49. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., Taigman, Y.: Text-to-4D dynamic scene generation. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 31915–31929. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/singer23a.html 2, 5
50. Sivakumar, P., Janson, P., Rajasegaran, J., Ambegoda, T.: Fewshotnerf: Meta-learning-based novel view synthesis for rapid scene-specific adaptation. arXiv preprint arXiv:2408.04803 (2024) 2
51. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 2, 3, 4
52. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021) 3
53. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In: Int. Conf. Learn. Represent. (2024) 2
54. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022) 1, 3
55. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=SJ1kSyO2jwu 1, 3
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 5

57. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual descriptions. In: International Conference on Learning Representations (2022) 1
58. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.00381, http://dx.doi.org/10.1109/CVPR52688.2022.00381 2
59. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023) 1, 2, 7
60. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36** (2023) 1, 2
61. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.02003, http://dx.doi.org/10.1109/CVPR52729.2023.02003 2
62. Zamani, A., Aghdam, A.G., Popa, T., Belilovsky, E.: Temporally consistent object editing in videos using extended attention. CVPR Workshop on AI for Content Creation (2024) 2
63. Zheng, Y., Li, X., Nagano, K., Liu, S., Hilliges, O., De Mello, S.: A unified approach for text- and image-guided 4d scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7300–7309 (June 2024) 2