CLIP-Actor: Text-Driven Recommendation and Stylization for Animating Human Meshes
— Supplementary Material —

Kim Youwang[1*] Kim Ji-Yeon[2*] Tae-Hyun Oh[1,2†]

Department of [1]EE & [2]CiTE, POSTECH
{youwang.kim, jiyeon.kim, taehyun}@postech.ac.kr
https://clip-actor.github.io

This supplementary material aims to provide additional contents and details that are not included in the main paper due to the space limitation. In Sec. A, we describe the design choices for the proposed Decoupled Neural Style Field (DNSF) and justify our full model by comparing it with different settings. In Sec. B, we describe the datasets and experimental details of human motion recommendation. In Sec. C, we provide additional qualitative results of CLIP-Actor. In Sec. D, we provide the overall algorithm for CLIP-Actor. Furthermore, we provide training details of CLIP-Actor in Sec. E and provide the discussion and possible future direction of CLIP-Actor in Sec. F. We also provide the video results demonstrating the text-conforming stylized meshes in motion.

A Analysis on the Decoupled Neural Style Field

In this section, we introduce our design choices and implementation details when composing and optimizing our decoupled neural style field, DNSF, and corresponding effects. Moreover, we provide implementation details of our mask-weighted embedding attention.

A.1 Effects of Content Mesh Resolution

Recall that the CLIP-Actor learns the best text-conforming per-vertex color and displacement for the content meshes. Thus, the resolution of the content meshes, *i.e.*, the number of mesh vertices, would be the critical factor of the texture generation quality. Text2Mesh also shows that the naïve neural style field network can synthesize a more plausible style with high-resolution meshes, *i.e.*, meshes with more vertices [8].

SMPL vs. SMPL-X: content mesh selection. We can change the content mesh model with SMPL [5] variants, SMPL-H [13] and SMPL-X [9], which have different numbers of vertices, to investigate the effects of the mesh resolution. With its linear blend skinning operation, SMPL maps pose parameters, $\mathbf{R}_t$ and shape parameters $\boldsymbol{\beta}_t$, to 6,890 mesh vertices, *i.e.*, $\mathbf{M}_{\text{SMPL}} \in \mathbb{R}^{6890\times 3}$. SMPL-X, on the other hand, has 10,475 vertices *i.e.*, $\mathbf{M}_{\text{SMPLX}} \in \mathbb{R}^{10475\times 3}$. Furthermore, SMPL-X can express detailed hand poses and expressive faces, which is essential

* Authors contributed equally to this work.
† Joint affiliated with Yonsei University, Korea.

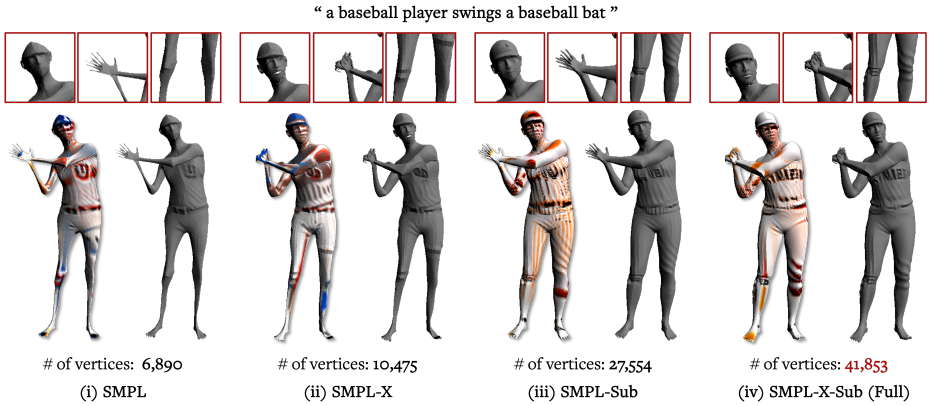

Fig. a. Effects of content mesh resolution. (i) With SMPL, which has the smallest number of vertices, CLIP-Actor shows unrealistic texture and geometric details. **(ii)** With higher resolution mesh, SMPL-X, CLIP-Actor achieves much smoother geometry, along with expressive hand and facial details. However, it still suffers from unrealistic colors. **(iii)** With subdivided SMPL, CLIP-Actor achieves better texture and geometry details than (i) and (ii). **(iv)** CLIP-Actor with subdivided SMPL-X achieves the most realistic color configuration and fine-grained geometric details.

in modeling human interactions and expressions [9]. We conduct experiments that compare the qualitative mesh stylization results with both mesh models as the content mesh. Figure a(i),(ii) illustrate qualitative results of CLIP-Actor with different body mesh models. With the lowest mesh resolution, *i.e.*, SMPL, CLIP-Actor generates unrealistic body configurations, such as sharp edges and slim body parts. Also, using SMPL as the content mesh, CLIP-Actor cannot represent detailed human actions. In Fig. a(i), the SMPL mesh spreads its hands, thus fails to express the baseball player grabbing the bat. On the other hand, SMPL-X, which has a higher resolution than SMPL, shows a smoother result, detailed hand pose, and facial expressions.

Mesh subdivision for higher resolution. Furthermore, we utilize mesh subdivision [12] to achieve higher mesh resolution ($\sim 4\times$ number of vertices). Note that we use subdivided SMPL-X for the content mesh for our full model[1].

Using subdivided meshes of SMPL and SMPL-X (Fig. a(iii),(iv)) as the content mesh, they show detailed cloth geometry, texture generation and smooth body curvatures than the basic SMPL, SMPL-X meshes. Also, SMPL-X-Sub, which is our CLIP-Actor's full version, shows the most realistic color configuration compared to other mesh models. Since our text-driven optimization of decoupled neural style field is based on the rendering of the stylized meshes, we postulate that higher resolution of the content meshes results in better supervision signal, thus leading to improved qualitative results.

[1] Note that we denote our content mesh as SMPL in the main paper for simplicity.

A.2 Mask-weighted Embedding Attention

In the main paper, we mentioned that 2D augmentations are essential for plausible texture generation. Recall that, we apply differentiable 2D augmentations before the rendered images are passed into the pre-trained CLIP encoder.

In practice, we adopt the multi-level 2D augmentations for the rendered images, following Text2Mesh [8]. The multi-level 2D augmentation is the method that renders both colored mesh and de-colorized mesh into images $\mathbf{I}^*$ and $\mathbf{I}^*_{geo}$, computes semantic loss for each rendered image, and leverage gradient accumulation during optimization. The advantages of such multi-level 2D augmentations are in two-folds. First, rendered images in diverse viewpoints and augmentations improve generalization across views [4]. Next, separate rendering of textured and de-colorized meshes, $\mathbf{I}^*$ and $\mathbf{I}^*_{geo}$, and gradient accumulation enable guiding both global context and local geometric details with only a single text prompt [8].

In detail, we apply a global 2D augmentation $\mathcal{T}_{global}(\cdot)$ to the rendered images $\mathbf{I}^*$. $\mathcal{T}_{global}(\cdot)$ does not contain the image crop but only random perspective transformation. Also, the local 2D augmentation $\mathcal{T}_{local}(\cdot)$ is applied to $\mathbf{I}^*$ and $\mathbf{I}^*_{geo}$. $\mathcal{T}_{local}(\cdot)$ contains both random crops up to 10% of the original image and the random perspective transformation.

However, the problem occurs in careless $\mathcal{T}_{local}(\cdot)$. Prior work [8] simply applied extreme close-ups to the de-colorized rendering of the meshes, which leads to random, empty rendered images. Such empty images do not conform to the text prompt, and these dummy images can distract the optimization process with random gradient direction. In CLIP-Actor, we mitigate this problem with mask-weighted attention embedding.

B Datasets of Human Motion Recommendation

In this section, we explain the dataset used in the retrieval system and evaluation for human motion recommendation. Moreover, we provide the details of the dataset and the experiment settings.

Retrieval dataset. We use BABEL [10] as a database of the retrieval system. BABEL is a dataset that labels a large-scale human motion capture dataset [6] with unique action categories. Although they provide over 250 action categories, *e.g.*, arm movements, the categories are too abstracted to be matched with our natural language prompt. Therefore, we utilize the raw labels untrimmed and diverse, *e.g.*, walk without energy and walk fast, so that the variants of natural language text prompt can be semantically matched with the raw labels. Instead of using a limited number of closed-set action categories, the raw labels can handle the open-set action descriptions.

Evaluation dataset. To evaluate the text-driven motion retrieval module, we use the SICK [7] as an evaluation dataset. SICK consists of the sentence pairs obtained from the Flickr8K dataset [3] and the video description dataset [1]. Since the sentences in SICK are composed of descriptions of images or video, the dataset is well-matched with our multi-modal retrieval scenario in terms of finding visual semantics. Each sentence pair in SICK is annotated with a

relatedness score from 1 to 5 that indicates the degree of semantic relatedness between two sentences. We set a range of scores from 4.4 to 4.8 for evaluation settings to ignore unreliable pairs and exclude the pairs that are only different with grammatical voice or article. SICK4.8 and SICK4.4 settings are constructed with the sentences with the score 4.8 and 4.4, respectively and SICK[4.4,4.8] setting comprises the sentences in the range. The samples of the SICK according to the score are shown in Table 1b in the main paper.

C Additional Qualitative Results

In this section, we present additional qualitative results of CLIP-Actor, with diverse subjects and actions (See Fig. b) We describe the text prompts we used and the corresponding results. Since we cannot express dynamic action sequences in images, video results are also attached in the supplementary files.

Fig. b. Additional qualitative results. The first two rows show that CLIP-Actor recommends the human motion from the text description and generates plausible mesh stylization in zero-shot. The last row shows the compositional mesh stylization that allows users to stylize the same retrieved motion, *"walking forwards"* with different identities via text prompt, *e.g.*, Hermione Granger.

D CLIP-Actor: Algorithm

In this section, we provide a thorough algorithm for CLIP-Actor. CLIP-Actor is a system that includes the recommendation module, text-driven DNSF optimization, and stylization, and we arrange the overall pipeline with the algorithm.

Algorithm 1 Overall pipeline of CLIP-Actor

Require: Pre-trained CLIP image encoder $\mathbf{g}(\cdot)$, text encoder $\mathbf{h}(\cdot)$, Pre-trained MPNet text encoder, $\mathbf{m}(\cdot)$, SMPL Linear Blend Skinning $\mathcal{M}(\cdot)$, BABEL dataset $\mathcal{A}$, SMPL template mesh $\mathbf{M}_c$

Input: Natural language text prompt y

Output: Text-conforming stylized meshes in motion $\mathbf{M}^*_{1:T}$

Text-driven Human Motion Recommendation

$\mathcal{S}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$ ▷ Cosine similarity between two vectors, $\mathbf{x}$, $\mathbf{y}$

`top-k`$[\cdot]$ returns `top-k` items and indices in tuple

$[\,\mathcal{A}_\text{k}, \_\,] \leftarrow$ `top-k`$[\mathcal{S}(\mathbf{h}(a_i), \mathbf{h}(y))], \;\; \forall a_i \in \mathcal{A}$ ▷ Cross-modal aware matching

$a_* \leftarrow \arg\max_{a_j \in \mathcal{A}_\text{k}} \mathcal{S}(\mathbf{m}(a_j), \mathbf{m}(y))$; ▷ Textual semantic matching

Get pose parameters from BABEL dataset with retrieved action label, a_*

$\mathbf{R}_{1:T} = [\mathbf{R}_1, \ldots, \mathbf{R}_T] \leftarrow$ `BABEL`(a_*)

$\mathbf{M}_{1:T} = \mathcal{M}(\mathbf{R}_{1:T}, \boldsymbol{\beta})$; ▷ Motion sequence of the content meshes

$\mathbf{I}_{1:T} \leftarrow$ `render`$(\mathbf{M}_{1:T})$;

$[\,\_, \text{idx}\,] \leftarrow$ `top-k`$[\mathcal{S}(\mathbf{g}(\mathbf{I}_{1:T}), \mathbf{h}(y))]$; ▷ Multi-modal content mesh sampling

DNSF optimization for L iterations

for iter $= 1, 2, \ldots, L$ **do**

  $\mathcal{L}_s \leftarrow 0$

  $\mathbf{c}, \mathbf{d} \leftarrow G_\theta(\mathbf{M}_c)$; ▷ Decoupled Neural Style Field

  for $i \in$ idx **do** ▷ Temporal view augmentation

    $\mathbf{M}^*_i \leftarrow$ `texturize`$(\mathbf{c}, \mathbf{d})$;

    Sample N camera poses, $\mathbf{p} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]$ ▷ 3D Spatial augmentation

    for $j = 1, 2, \ldots, N$ **do**

      $\mathbf{I}^*_{ij} \leftarrow$ `render`$(\mathbf{M}^*_i, \mathbf{p}_j)$;

      $\mathbf{I}^*_{ij} \leftarrow$ `2D_augmentations`$(\mathbf{I}^*_{ij})$; ▷ 2D Spatial augmentation

      $w_{ij} \leftarrow$ `mask_weighted_att`$(\mathbf{I}^*_{ij})$;

    end for

    $\bar{\mathbf{g}}(\mathbf{I}^*_i) = \frac{\sum_{j=1}^N w_{ij} \mathbf{g}(\mathbf{I}^*_{ij})}{\sum_{j=1}^N w_{ij}}$; ▷ Mask-weighted embedding attention

    $\mathcal{L}_s \leftarrow \mathcal{L}_s + (1 - \mathcal{S}(\bar{\mathbf{g}}(\mathbf{I}^*_i), \mathbf{h}(y)))$;

  end for

  $\theta^* \leftarrow$ Update DNSF G_θ parameters, θ

end for

Test time: Stylization of human meshes in motion

$\mathbf{c}^*, \mathbf{d}^* \leftarrow G_{\theta^*}(\mathbf{M}_c)$ ▷ Generate color and geometry with learned DNSF

for $k = 1, 2, \ldots, T$ **do**

  $\mathbf{M}^*_{1:T} \leftarrow$ `texturize`$(\mathbf{c}^*, \mathbf{d}^*)$ ▷ Stylize meshes in motion with $\mathbf{c}^*, \mathbf{d}^*$

end for

E Training Details

We provide training details of CLIP-Actor, including optimizer, training hardware specifications, and training time. We use the Adam optimizer with the initial learning rate set to 0.0005 and the learning rate decay factor as 0.9 every 100 iterations. We train CLIP-Actor for 1500 iterations using a single NVIDIA TITAN RTX GPU. Total training takes about 30 minutes to one hour depending on the motion sequence length, and training options such as the number of frames we use for spatio-temporal view augmentation.

F Discussion

We find the observations about human mesh stylization harnessing CLIP [11] text-image joint space. Given the text prompt that describes an interaction with objects, the objects are often projected onto the human mesh and stylized together. For example, a basketball is depicted on the player's chest when the action prompt that interacts with the ball is given, *i.e.*, chest passing (see Fig. c). Since CLIP is trained with the pairs of text and 2D images, a depth ambiguity from the 2D images can be propagated to the 3D mesh stylization. The further development of object mesh manipulation can be applied to our work to model Human-Object-Interaction [2, 14] as future work.

Fig. c. A case of object projection on mesh surface.

Since CLIP-Actor recommends the motion conforming to the input prompt instead of generating motions, some prompts might not be compatible with the BABEL [10]. CLIP-Actor has two major features to prevent such cases. First, our retrieval module implements semantic matching; thus, it finds visual and textual proximal action labels robustly. For example, given "Thor swinging Mjölnir" as an input, where "swinging Mjölnir" is not included in BABEL, CLIP-Actor retrieves "swing hammer side to side."

Still, incompatible prompts might exist and harm the subsequent stylization process. Our multi-modal content mesh sampling handles such cases. It finds the best mesh frames within the motion to achieve reasonable stylization quality. Our design choices on modules prevent the drastic degradation in stylization even with incompatible prompts. We think that further improvements to handle out-of-distribution cases would be an interesting future direction.

References

1. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: Semeval-2012 task 6: A pilot on semantic textual similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. p. 385–393 (2012) 3
2. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. IEEE International Conference on Computer Vision (ICCV) (2021) 6
3. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013) 3
4. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 3
5. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (SIGGRAPH Asia) **34**(6), 248 (2015) 1
6. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: IEEE International Conference on Computer Vision (ICCV) (2019) 3
7. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A sick cure for the evaluation of compositional distributional semantic models. In: International Conference on Language Resources and Evaluation (LREC) (2014) 3
8. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 1, 3
9. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 2
10. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3, 6
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021) 6
12. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. arXiv:2007.08501 (2020) 2
13. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (SIGGRAPH Asia) **36**(6) (Nov 2017) 1
14. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision (ECCV) (2020) 6