You Only Touch Once: One-Touch System for Personalized 3D Music Video Generation

Kyungjune Lee*‡, Youngjin Shin†‡, Jungwoo Huh*, and Sanghoon Lee*§
* Yonsei University, Seoul 03722, South Korea
E-mail: {naive2kj90, gjwjddn9, slee}@yonsei.ac.kr
† Yonsei University, Seoul 03722, South Korea
E-mail: shinstar1214@naver.com

***Abstract*—Recent advances in generative artificial intelligence (AI) have enabled the synthesis of multi-modal content, including dance motion, camera trajectories, and rendered 3D scenes. However, most existing approaches rely on fragmented pipelines that require significant manual effort and domain expertise. In this paper, we propose an end-to-end, one-touch system for generating fully rendered 3D music videos, aimed at non-expert users. Our proposed system integrates 3D avatar generation, motion transfer, camera trajectory generation, and scene rendering, all built upon a unified dataset captured via synchronized multi-view imaging and motion capture. To evaluate the quality of the generated videos, we construct a multi-modal dataset that jointly aligns motion, music, and camera trajectory. Experimental results demonstrate high perceptual scores in synchronization and visual coherence across user groups, with a low incidence of rendering artifacts. Our proposed system significantly lowers the barrier to creating high-quality 3D music videos and offers an extensible platform for multi-modal content generation. The demonstration video is available at https://youtu.be/FDQ0peWFSio.**

I. Introduction

Recent advancements in multi-modal generative models have enabled users to create personalized content across multiple domains, including music, motion, and camera trajectories. In particular, models for music-to-dance generation [1]–[3] synthesize plausible motion from music, while recent work on camera trajectory generation [4]–[7] produces dynamic camera trajectory conditioned on music or textual descriptions.

While these cross-modal models have begun to blur the boundaries between previously separate content types, practical content creation pipelines remain highly fragmented. From avatar construction and motion transfer to camera design and final rendering, existing approaches rely on heterogeneous tools and ad hoc integration. These pipelines demand significant domain knowledge, limiting their accessibility for non-expert users.

Moreover, transferring motion data to 3D avatars is non-trivial due to inconsistencies in joint template definitions across datasets and models (e.g., [8] uses 51 joints, [4] 61 joints, and [9] 24 joints). Similarly, camera parameters generated by different models often vary in format and coordinate convention, requiring manual post-processing to ensure interoperability. These technical inconsistencies pose substantial barriers to scalable content generation.

‡ Equal contribution
§ Corresponding author

(a) 3D Avatar variation (b) Motion variation (c) Background variation

Fig. 1. The results from YOTO showing (a) the variation in background using the same dance motion and 3D avatar, (b) the variation in motion which results from different camera trajectories, and (c) the variation in the 3D avatar with different subjects.

To address these limitations, we propose a one-touch system, **You Only Touch Once (YOTO)**, for generating fully rendered 3D music videos. Here, "one-touch" refers to a system that operates with a single user action—once the user provides input, all subsequent stages. The proposed system integrates four key components—3D avatar generation, motion transfer, camera trajectory synthesis, and scene rendering—into a single cohesive pipeline.

Also, we construct a new multi-modal dataset that provides synchronized sequences of music, motion, camera trajectories, and backgrounds. Existing datasets typically focus on only a subset of these modalities, making it difficult to evaluate cross-domain coherence in generated content. Our dataset addresses this gap and enables end-to-end generation within a unified pipeline.

Finally, we conduct both quantitative and qualitative evaluations. Quantitative analysis using Mean Per Joint Position Error (MPJPE) confirms that the generated motion retains structural fidelity under rendered views, achieving a mean MPJPE of 39.2 mm. Qualitative results, based on user studies across motion–music synchronization and visual harmony, further validate the perceptual realism of our system. Participants rated the generated videos with a score of 3.90 for synchronization accuracy and 3.75 for visual harmony on a 5-point Likert scale, indicating strong perceptual quality and coherence. These findings demonstrate the feasibility of accessible, high-quality, cross-domain video generation without requiring expert knowledge.

Our contributions are as follows:

979-8-3315-7206-8/25/$31.00

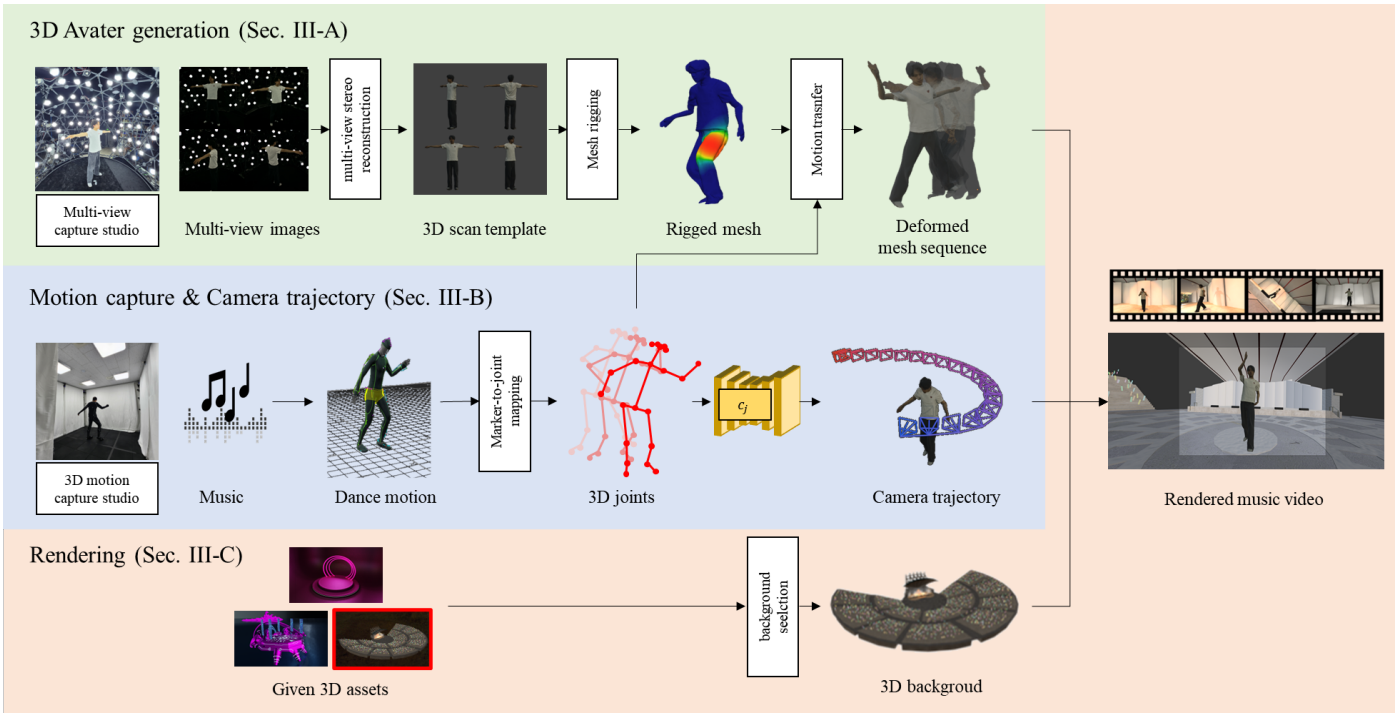

Fig. 2. An overview of YOTO. We first construct an animatable 3D avatar by combining a multi-view reconstructed mesh with motion capture data. Then, we synthesize a complete music video by integrating avatar motion, music-conditioned camera trajectories, and user-selected background within a unified rendering framework.

1) We present an end-to-end one-touch system, YOTO, that generates fully rendered 3D music videos by unifying avatar construction, motion transfer, camera trajectory synthesis, and scene rendering. Our system removes the need for expert intervention by integrating these heterogeneous components into a streamlined, user-friendly pipeline.
2) Our system enables flexible coupling between modalities such as music, dance motion, and camera trajectory. Users can dynamically generate diverse video outputs that reflect varying combinations of content elements, supporting personalized content creation.
3) We construct a new dataset comprising synchronized music, motion, camera trajectories, and background scenes. This comprehensive alignment across modalities facilitates both training and evaluation of cross-domain generative systems, which was previously impractical with existing datasets.

II. System

Our YOTO leverages both a multi-view capture studio and a motion capture studio. From a multi-view capture studio, we acquire 360-degree RGB images of a subject and reconstruct a 3D mesh of RGB images, which becomes a 3D Avatar of a subject The motion capture studio utilizes markers to track the subject's movement, providing joint-level position and rotation data of each 51 joints.

a) Multi-view capture studio: We construct a multi-view capture studio, as shown in the green box of Fig. 2. The studio employs 30 RGB vision cameras with a resolution of 1024×1024 at 30 frames per second (fps). The cameras are arranged in a dome-like configuration to enclose the target object. All cameras are synchronized using external trigger devices (*KOTRON TG-16C* and *KOTRON TG-4C*), following the multi-sensor synchronization protocol introduced in [10]. The primary trigger, *KOTRON TG-16C*, generates periodic synchronization signals, while two sub-triggers, *KOTRON TG-4C*, operate in bypass mode to distribute signals to 10 and 11 cameras, respectively. To support parallel data acquisition, five desktop computers running *Microsoft Windows 10* are deployed, with each desktop responsible for six cameras based on bandwidth constraints. Another desktop is designated as the master host and coordinates all camera operations through TCP/IP network communication.

b) 3D motion capture studio: For high-precision human motion acquisition, we employ a dedicated OptiTrack system [8] consisting of six infrared motion capture cameras as shown in the blue box of Fig. 2. The cameras are mounted at the top corners and lateral upper edges of a 3.3 m × 3.3 m × 3.0 m studio space, providing a wide field of view with minimal occlusions. Each camera captures at 120 fps, enabling accurate tracking of fast human motion from markers. The acquired motion data comprises 51 joints, including 21 body joints and 30 hand joints. Motion data and music at frame t, each denoted as θ_t and m_t, are synchronized and collected in real-time.

To capture expressive dance motion, subjects are instructed to perform freestyle dance in synchronize with a given music

during recording. This ensures temporal alignment between the captured motion data and the accompanying music.

III. Method

As shown in Fig. 2, our proposed YOTO initiates by capturing the user's 3D avatar and dance motion synchronized to music. Then, the captured dance motion is combined with the 3D avatar through rigging to produce an animatable 3D avatar. Subsequently, a camera trajectory is generated and applied to the dancing 3D avatar within a 3D background. This entire sequence is rendered along with the original music, as shown in the red box of Fig. 2, resulting in a complete 3D music video through our one-touch generation pipeline. The subsequent sections provide detailed descriptions of each component of the YOTO.

A. 3D avatar generation

To obtain a realistic 3D body scan, we capture a set of static multi-view images and perform reconstruction in a coarse-to-fine pipeline. Initially, coarse mesh geometry $\mathbf{v}_c$ is estimated using multi-view stereo (MVS) techniques [11], followed by Poisson surface reconstruction to generate a watertight mesh. To recover fine-scale details, we optimize a per-vertex displacement field $\Delta\mathbf{v}$, resulting in the refined body mesh $\mathbf{v}_b = \mathbf{v}_c + \Delta\mathbf{v}$. The optimization is jointly performed over the displacement field $\Delta\mathbf{v}$ and the coefficients of spherical harmonics lighting. Our objective is to minimize the discrepancy between the rendered shading of the refined mesh and the grayscale appearance of the input images. This is motivated by the observation that grayscale intensities serve as a reasonable approximation of surface shading. The total energy function is defined as:

$$E = E_s + \lambda_e E_e + \lambda_n E_n + \lambda_m E_m, \tag{1}$$

where E_s denotes the shading reconstruction error, and E_e, E_n, and E_m represent regularization terms encouraging edge length preservation, normal consistency, and Laplacian smoothness, respectively. We empirically set the weighting parameters to $\lambda_e = 0.01$, $\lambda_n = 0.1$, and $\lambda_m = 0.1$. Fig. C presents qualitative comparisons with template-based and direct reconstruction approaches. Our method achieves noticeably improved geometric fidelity, capturing fine mesh details more accurately than baseline methods.

B. Motion capture and camera trajectory

To synthesize motion into a 3D avatar, we transfer captured motion data into a rigged mesh in initial pose V_0 with motion transfer function F, which applies joint rotation to the 3D avatar. Animated avatar at frame t, V_t, can be written as,

$$V_t = F(V_0, \theta_t). \tag{2}$$

Also, we generate camera trajectory $C_t \in \mathbb{R}^8$ at frame t. The camera trajectory is represented with position and orientation parameters as follows:

$$C_t = [p_t^x, p_t^y, p_t^z, r_t^x, r_t^y, r_t^z, d_t, \phi_t], \tag{3}$$

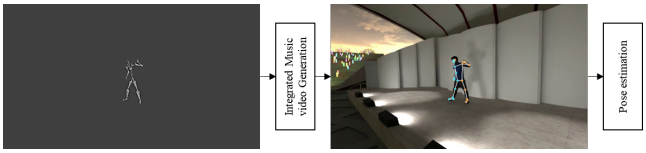

Fig. 3. Structural consistency evaluation between the original motion and the rendered video. We estimate pose from rendered frames and compare them with the original captured motion data to assess motion preservation under varying camera and background conditions.

where (p_t^x, p_t^y, p_t^z) denote the gloabl position of the camera, (r_t^x, r_t^y, r_t^z) represent the Euler rotation angles, d_t is the distance to the target, and ϕ_t is the field of view (FOV) at frame t. The camera trajectory C_t is generated using a cross-modal model conditioned on motion and music features:

$$C_t = G(\theta_t, m_t), \tag{4}$$

where G is a state-of-the-art camera trajectory generation model [4].

C. Rendering

To accommodate diverse user preferences and enable flexible scene composition, the background BG is designed to be customizable within the rendering pipeline. A set of 3D background scenes with varying visual styles is provided, allowing users to select one that best matches the intended atmosphere. At each time frame t, a rendered image I_t from integrating the posed 3D avatar, camera parameters, music features, and selected background can be written as,

$$I_t = R(V_t, C_t, m_t, BG), \tag{5}$$

where R is the rendering function. The complete video is then constructed by temporally concatenating the sequence of rendered frames $\{I_t\}_{t=1}^{T}$.

IV. Experiments

A. Dataset Construction

To evaluate the alignment quality among motion, camera, and music, we construct a dataset of 33 synchronized data pairs. Each pair includes a professional dancer's motion, music, and camera trajectory. The motion is captured in a studio as described in Sec. III-B and applied to 3D avatars generated in Sec. III-A.

Unlike existing datasets, ours integrates music, motion, camera, and background in a tightly aligned manner. Comparison of our dataset with prior works based on these four components are shown in Table I.

B. Quantitative Evaluation

Well-generated video should faithfully preserve the structure and timing of the original motion data, ensuring that camera movement and background enhance rather than interfere with the viewing experience. To evaluate this, we conduct joint-level tracking between the original motion data at frame t, denoted as J_t^i for the i-th joint, and the detected joints $\tilde{J}_t^i$ from the final

TABLE I
COMPARISON OF DATASETS BY MUSIC, DANCE MOTION, CAMERA TRAJECTORY, AND BACKGROUND.

Dataset	Music	Dance Motion	Camera Trajectory	Background	Images	Subjects	Sequences
AIST++ [12]	◯	◯	△	×	◯	30	1,408
DanceCamAnimator [4]	◯	◯	◯	×	×	-	108
BABEL [13]	×	△	×	×	×	300	11k
UBody [14]	×	△	△	×	◯	-	-
Ours	◯	◯	◯	◯	◯	**10**	**33**

rendered music videos using the same view C_t, by leveraging a pose estimation model [§]. To quantify the positional accuracy of the synthesized motion, we compute the mean per joint position error (MPJPE) as follows:

$$\text{MPJPE} = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{i=1}^{N} \left\| J_t^i - \tilde{J}_t^i \right\|_2, \tag{6}$$

where T is the total number of frames, N is the number of joints, J_t^i is the ground-truth position of the i-th joint at frame t, and $\tilde{J}_t^i$ is the corresponding predicted position from the rendered video. Frames with missing detections from [15] are excluded from the computation.

Our YOTO achieves mean and median MPJPE as **39.2 mm**, **39.9 mm**, respectively. This value indicates that the joint positions estimated from rendered videos remain closely aligned with the original captured motion data, even after rendering and camera application. Our results fall within the range typically reported by recent 3D human motion generation methods [16]–[18]. These support the idea that our YOTO produces structurally faithful, temporally consistent motions under dynamic camera views, satisfying widely accepted criteria for well-reconstructed motion.

C. Qualitative Evaluation

To assess the perceptual quality of the generated videos, we conduct a user study using 700 short video clips (5 s each, 150 frames), covering various combinations of motion, music, camera, and background. A total of 60 participants, including 10 professional dancers, 10 graphic design experts, and 40 general users, participate in the evaluation.

Each participant rates the videos across two perceptual dimensions: (a) synchronization accuracy, representing the alignment between dance motion and music rhythm, and (b) visual harmony, assessing the coherence among avatar, motion, camera, and background. Ratings are given on a 1–5 Likert scale, where 5 indicates the highest quality. As shown in Fig. 4, participants have rated our YOTO favorably across all groups. The dancer group has rated 4.10 for synchronization accuracy and 3.80 for visual harmony, while the designer group has rated it 3.40 and 3.20, respectively. The general user group has rated scores of 3.98 for synchronization and 3.88 for visual harmony. On average, our YOTO has scored 3.90 for synchronization accuracy and 3.75 for visual harmony. These results indicate that the YOTO produces perceptually consistent outputs across user groups with varying domain expertise.

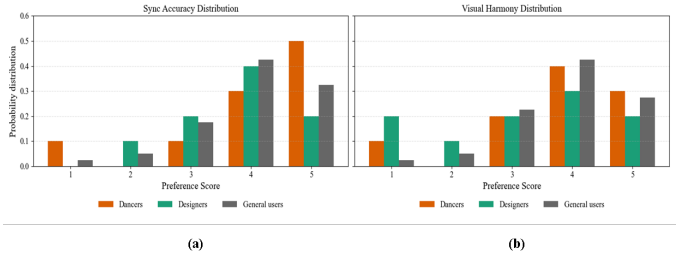

Fig. 4. Perceptual evaluation results. (a) Synchronization accuracy between dance motion and music rhythm. (b) Visual harmony among the avatar motion, camera trajectory, and background scene. Each is evaluated across three participant groups: professional dancers, designers, and general users.

V. CONCLUSION

In this paper, we present YOTO, a one-touch system for generating fully rendered 3D music videos by unifying 3D avatar generation, motion transfer, camera trajectory generation, and scene rendering into a single, cohesive pipeline. Our YOTO is designed to eliminate the fragmentation of existing multi-stage tools and to lower the barrier for non-expert users.

To support this framework, we constructed a tightly aligned dataset that integrates music, motion data, camera trajectory, and background. Built upon this dataset, our generation pipeline was evaluated both quantitatively and qualitatively. Results demonstrate that the generated content retains structural motion fidelity and achieves high perceptual quality under varied camera and background conditions.

By enabling modular customization while preserving temporal and spatial coherence, our system facilitates the scalable production of hyper-quality content. We believe this work represents a significant step toward democratizing 3D content creation and opens new directions for personalized AI-driven storytelling. Future extensions will explore real-time interaction, style adaptation, and broader avatar and scene diversity to further expand usability and creative flexibility.

[§] We adopt the pose estimation model proposed in [15], which estimates 3D human joint locations from RGB images as shown in Fig. 3. The model employs a transformer-based architecture and has demonstrated strong generalization on in-the-wild datasets. In our setup, we feed rendered video frames into the model to obtain $\tilde{J}_t^i$ for each joint i at each frame t, enabling direct comparison with the original joint locations J_t^i

ACKNOWLEDGMENT

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2024-00398413, Contribution Rate: 100%), and the Yonsei Signature Research Cluster Program of 2025 (2025-22-0013).

REFERENCES

[1] W. Zhuang, C. Wang, J. Chai, Y. Wang, M. Shao, and S. Xia, "Music2dance: Dancenet for music-driven dance generation," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2, pp. 1–21, 2022.

[2] J. Tseng, R. Castellon, and K. Liu, "Edge: Editable dance generation from music," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 448–458.

[3] S. Kim and K. Lee, "Music-driven synchronous dance generation considering k-pop musical and choreographical characteristics," *IEEE Access*, vol. 12, pp. 94 152–94 163, 2024. DOI: 10.1109/ACCESS.2024.3420433.

[4] Z. Wang, J. Li, X. Qin, *et al.*, "Dancecamanimator: Keyframe-based controllable 3d dance camera synthesis," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 200–10 209.

[5] H. Jiang, X. Wang, M. Christie, L. Liu, and B. Chen, "Cinematographic camera diffusion model," in *Computer Graphics Forum*, Wiley Online Library, vol. 43, 2024, e15055.

[6] R. Courant, N. Dufour, X. Wang, M. Christie, and V. Kalogeiton, "E.t. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness," in *European Conference on Computer Vision*, Springer, 2024, pp. 464–480.

[7] M. Zhang, T. Wu, J. Tan, Z. Liu, G. Wetzstein, and D. Lin, "Gendop: Auto-regressive camera trajectory generation as a director of photography," *arXiv preprint arXiv:2504.07083*, 2025.

[8] OptiTrack, *OptiTrack Motion Capture System*, https://optitrack.com.

[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[10] K. Lee, J. Lee, H. Lee, M. Jang, S. Lee, and S. Lee, "Faceclone: Interactive facial shape and motion cloning system using multi-view images," in *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, 2023, pp. 512–513.

[11] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2495–2504.

[12] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 401–13 412.

[13] A. R. Punnakkal, A. Chandrasekaran, N. Athanasiou, A. Quiros-Ramirez, and M. J. Black, "Babel: Bodies, action and behavior with english labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 722–731.

[14] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, "One-stage 3d whole-body mesh recovery with component aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 159–21 168.

[15] C. Lugaresi, J. Tang, H. Nash, *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. [Online]. Available: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf.

[16] Y. Li, Z. Huang, W. Xu, and P. Luo, "D3dp: Dual disentangled 3d pose learning from 2d videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[17] X. Chen, Y. Wang, J. Zhang, S. Zhu, and D. Lin, "Diffpose: Disentangled diffusion-based pose generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[18] H. Luo, X. Wang, H. Wang, Y. Wu, W. Zhang, and X. Li, "Stcformer: Spatio-temporal cross-interaction transformer for 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.