Text-Driven Artistic Staging: Pose, Lighting, and Camera References from Paintings

YUNGE WEN, Massachusetts Institute of Technology, Live Demo: https://artstage.darkmesh.art/

arXiv:2608.28823v1 [cs.CV] 28 Aug 2026

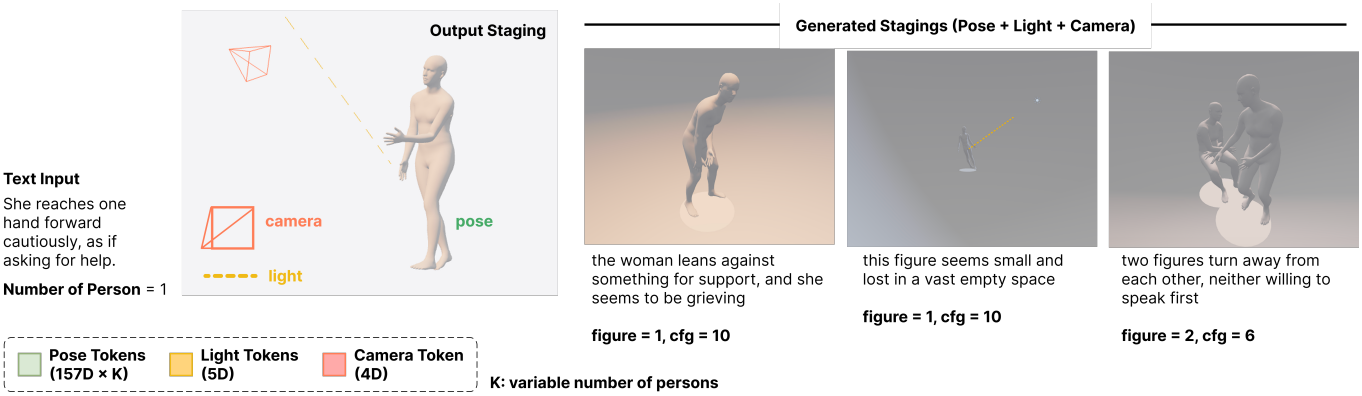

Fig. 1. A staging is a pose for every figure, a light and a camera, generated together from one sentence and editable afterwards. Left: the three components of a single sample, from a description that names the body explicitly. Right: samples for three affective descriptions of the kind ArtEmis annotators actually write, which name a feeling and leave the body and the framing to the model; one has two figures.

Artists coordinate human pose, illumination, and camera placement to convey narrative and emotion, but existing generative methods typically model these elements independently. We introduce text-to-editable 3D staging, a task that jointly generates human poses, a dominant light, and a camera configuration from an affective description. We construct 11,911 text–staging pairs from 2,328 figurative paintings by reconstructing SMPL bodies, estimating low-frequency illumination, recovering camera parameters, and pairing each scene with ArtEmis descriptions. We train a flow-matching transformer that supports variable numbers of figures and produces multiple staging alternatives for each prompt. On held-out descriptions, the model achieves 32.2% retrieval R@1, compared with 16.6% for CLIP-based nearest-neighbor retrieval, while approximately preserving corpus-level diversity. These results demonstrate the feasibility of generating editable, emotionally conditioned 3D staging references from text.

1 Introduction

Composition and layout are foundational to painting, storyboarding, theatre staging and photography alike. Staging an emotionally charged figure, a human artist intuitively settles three things at once: how the body is held, where the light falls, and where the viewer stands. A low viewpoint reads as looming only if the figure is upright; a rim light isolates only if the pose turns away. Choosing any one commits the other two.

Existing tools address one component at a time. Pose libraries and 3D mannequins model the body and leave lighting and framing to the artist. Generative methods are partitioned the same way: text-to-pose and text-to-motion work outputs the body alone [4, 11], camera trajectory methods [3, 6] take character motion as a conditioning *input* and do not model light, and relighting for images and video [7] produces images rather than a 3D scene. To our knowledge, existing methods do not jointly generate an editable staging from text describing the intended emotion or narrative.

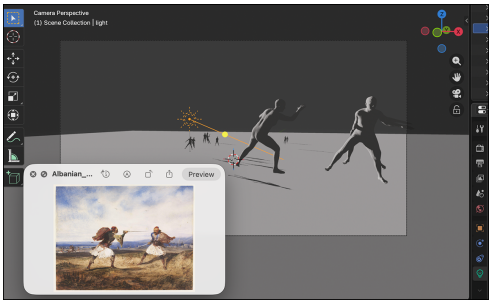

Fig. 2. A recovered staging beside the painting it came from. Every figure is an SMPL body, the light is a direction and a strength fitted to the heads, and the camera is the reconstruction viewpoint; the result is a scene an artist can open and edit.

We recover the joint solution from figurative painting, where pose, lighting and camera were resolved together to make a feeling legible. We contribute (i) a corpus of 11,911 (emotional text, 3D staging) pairs reconstructed from paintings and paired with ArtEmis [1] annotator responses; (ii) a flow-matching transformer that generates pose, lighting and camera jointly from one sentence, giving several distinct alternatives per query; and (iii) an evaluation showing that guidance trades text alignment against variation.

2 Dataset

From the intersection of ArtEmis [1] and WikiArt, we retain paintings that serve as professionally composed staging examples associated with an emotion agreed upon by most annotators and whose style supports reliable human reconstruction. Mask R-CNN then discards paintings with no person, figures that are too small, or more than four figures, leaving 6,826 candidates.

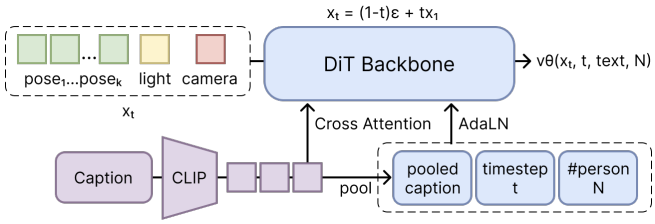

Fig. 3. Model Architecture

Pose. Each detected box goes to HMR2 [5] for an SMPL [9] body; the projected silhouette against the detector's mask scores the fit.

Lighting. Visible SMPL vertices within 0.11 m of the head centroid are sampled from the image, and least squares over the resulting (normal, luminance) pairs fits the first-order spherical harmonic luminance $= c_0 + \mathbf{c}_1 \cdot \mathbf{n}$. The direction and magnitude of $\mathbf{c}_1$ estimate the dominant low-frequency illumination and its directional variation. Per-channel fits estimate illumination chromaticity, from which we derive a correlated colour temperature.

Camera. Position follows from the reconstruction, with the group's lowest vertex on the ground. Field of view is fixed at 45°. For multiple figures, the 4DHumans crop-to-full conversion maps each reconstruction into a shared full-image camera coordinate system.

Filtering on these scores (Fig. 2) keeps 2,328 paintings, 34% of candidates, silhouette IoU and lighting fit accounting for almost all of the loss. Up to six ArtEmis sentences per work, each encoded by CLIP [10] into a token sequence, give $N = 11{,}911$ description–staging pairs.

3 Method

Representation. A staging is a sequence of typed vectors $[\, p_1 \ldots p_K,\ \ell,\ c \,]$ for $K \leq 4$ figures: $p_i \in \mathbb{R}^{157}$ is a body (24×6 joint rotations, 10 SMPL shape coefficients, 3 root position), $\ell \in \mathbb{R}^5$ the light (a unit direction as xyz, directional strength, colour temperature), $c \in \mathbb{R}^4$ the camera (elevation, azimuth as sine and cosine, log distance). Typed tokens let scenes with different figure counts share one model without padding, and let the figures attend to each other. The two quantities that wrap, camera azimuth and light direction, are stored as sine and cosine or as a unit vector; elevation is bounded and kept as an angle, and distance is stored as its logarithm. Sagittal reflection augments half of each epoch, mirroring the light and camera azimuth alongside the body.

Model. A pre-norm transformer of four layers at width 256 (Fig. 3). The CLIP-encoded description enters through cross-attention as a token sequence rather than pooled, so that the camera and light tokens can attend to the words describing them.

Objective. Flow matching [8], as in recent camera-control work [2]: with $x_t = (1-t)\,\epsilon + t\,x_1$ the network regresses $x_1 - \epsilon$, sampled in 30 Euler steps. Descriptions are dropped 10% of the time, making the guidance weight w available at sampling time. Training uses dropout 0.2, EMA and early stopping on a scene-level split.

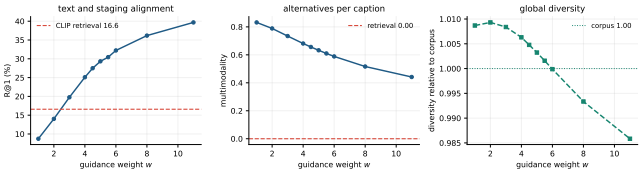

Fig. 4. Effect of guidance weight. Text alignment increases as w rises, while alternatives per description decrease. The model surpasses CLIP retrieval at $w \approx 2.5$ and approximately matches corpus-level diversity at $w = 6$. Retrieval has zero multimodality because it returns one staging per description.

4 Evaluation

Following text-conditioned camera generation [3, 6], we train and freeze a contrastive language–staging encoder. We measure alignment by R@1 over 32 held-out descriptions (3.125% chance), diversity across descriptions, and multimodality (MM) across repeated samples. As a direct test against corpus reuse, retrieval selects the staging whose training caption is nearest to the query in frozen CLIP space and has zero MM by construction.

We ablate classifier-free guidance in Fig. 4. Increasing w improves alignment but reduces MM. At $w = 6$, our model reaches 32.2% R@1 versus 16.6% for retrieval, retains 0.589 MM, and approximately matches corpus-level diversity.

5 Discussion and Future Work

Most ArtEmis descriptions express affect without specifying pose, lighting or framing, limiting direct supervision for each component. The corpus is dominated by portraits and therefore covers a narrow range of poses. Field of view is assumed rather than estimated, multiple figures can be placed closer together than in the source paintings, and the recovered lighting represents only an approximate low frequency cue.

Future work could weight training by lighting confidence and use film data to broaden pose and shot distributions. We also plan a study with professional illustrators comparing our stagings with pose libraries and image search.

References

[1] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas. 2021. ArtEmis: Affective Language for Visual Art. In *CVPR*.
[2] S. Bahmani, I. Skorokhodov, G. Qian, A. Siarohin, W. Menapace, A. Tagliasacchi, D. B. Lindell, and S. Tulyakov. 2025. AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers. In *CVPR*.
[3] R. Courant, N. Dufour, X. Wang, M. Christie, and V. Kalogeiton. 2024. E.T. the Exceptional Trajectories: Text-to-camera-trajectory generation with character awareness. In *ECCV*.
[4] Y. Feng, J. Lin, S. K. Dwivedi, Y. Sun, P. Patel, and M. J. Black. 2024. ChatPose: Chatting about 3D Human Pose. In *CVPR*.
[5] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *ICCV*.
[6] H. Jiang, X. Wang, M. Christie, L. Liu, and B. Chen. 2024. Cinematographic Camera Diffusion Model. *Computer Graphics Forum* 43, 2.
[7] P. Kocsis, J. Philip, K. Sunkavalli, M. Nießner, and Y. Hold-Geoffroy. 2024. LightIt: Illumination Modeling and Control for Diffusion Models. In *CVPR*.
[8] X. Liu, C. Gong, and Q. Liu. 2022. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv:2209.03003*.
[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* 34, 6.
[10] A. Radford, J. W. Kim, C. Hallacy, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.
[11] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. 2023. Human Motion Diffusion Model. In *ICLR*.