MoZoo: Bringing Animal Meshes to Life with Multimodal Video Diffusion

Anonymous authors
Paper under double-blind review



Figure 1: Given a mesh video, MoZoo synthesizes a photorealistic animal video using appearance evidence from text, an image, or a temporally independent reference video.

Abstract

Creating cinematic-quality animal effects is highly valuable but remains labor-intensive, as producing realistic animal appearance and motion still relies heavily on manual animation. Recent advances in video generation offer a promising path toward automating this process, yet existing models provide limited structural control over animal motion and geometry. Progress toward such controllable generation is further hindered by the lack of large-scale data pairing animated meshes with realistic RGB videos. We introduce **mesh-guided animal appearance synthesis**, a new formulation in which an animated mesh specifies geometry and motion, while a text, image, or video reference specifies appearance. To enable learning under this formulation, we construct **MoZoo-Data**, a 62K-clip synthetic-to-real dataset that combines aligned rendered pairs with filtered pseudo-pairs recovered from real videos via inverse mesh generation. Building on this data, we propose **MoZoo**, an in-context video diffusion model that synthesizes realistic animal videos while following mesh-defined motion and preserving reference appearance. MoZoo introduces **RAR-RoPE** to establish correspondence between mesh and target representations while handling temporally asynchronous references, and **ADA** to incorporate structural and appearance conditions without allowing noisy target features to contaminate the conditioning streams. We further introduce **MoZooBench**, a 120-case benchmark for evaluating structural consistency and appearance fidelity. Experiments show that MoZoo outperforms baselines trained on the same data and generalizes across reference modalities, geometries, and motion patterns.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Animal characters are integral to films and virtual worlds, where their realism shapes world-building and storytelling. Creating lifelike digital animals requires modeling anatomy, non-rigid deformation, and fur dynamics. As illustrated in Fig. 2 (top), conventional workflows combine skeletal rigging and animation, muscle simulation, and fine-grained fur grooming. These stages require experienced artists and technical directors to coordinate multiple physical systems and refine subtle visual details. The resulting production cost makes diverse, high-quality animal content difficult to scale.

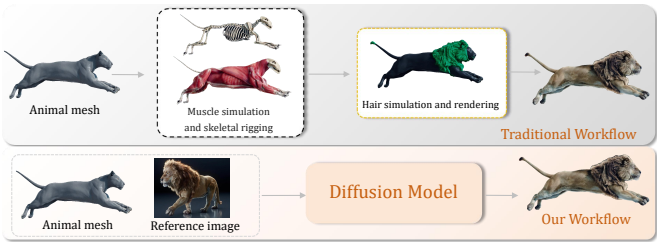

Figure 2: Conventional workflow (top) and MoZoo (bottom). The mesh sets geometry and motion; a text, image, or video reference sets appearance.

Conditional video generation offers a route to rendering existing animation assets. General video models provide text- and image-guided synthesis Wan et al. (2025); HaCohen et al. (2024); Yang et al. (2024). Recent task-specific methods address sketch-guided cartoon video synthesis Li et al. (2025a), human face swapping Guo et al. (2026), camera planning and human-scene animation Li et al. (2025b); Liu et al. (2026), and video-driven mesh animation Shi et al. (2025). These methods address different input-output settings from ours: we generate a realistic animal video from a given mesh animation and a separate text, image, or video appearance reference. This requires fidelity to mesh-defined motion and faithful rendering of diverse animal appearances.

We formulate mesh-guided animal appearance synthesis and introduce **MoZoo**, a video diffusion framework. As shown in Fig. 2(bottom), given a rendered video of an animated mesh and a text, image, or video appearance reference, MoZoo generates a realistic animal video that follows the mesh-defined geometry and motion while matching the reference appearance. This task presents two challenges. First, building large-scale, accurately aligned mesh–RGB video pairs that capture real animal appearances is difficult. Second, the model must align target frames with mesh guidance while incorporating temporally asynchronous appearance references without interference between structural and appearance conditions.

To address the data bottleneck, we construct **MoZoo-Data**, a synthetic-to-real dataset containing 62K clips for mesh-guided animal appearance synthesis. We use Unreal Engine 5 (UE5) to animate and render publicly available animal character mesh assets, generating paired coarse mesh sequences and RGB videos with accurate structural correspondence. However, synthetic data exhibits a domain gap from real animal videos, limiting models' ability to learn diverse real-world appearances. To expand real-world data, we train an inverse generative model on synthetic paired data and apply it to real animal videos to estimate corresponding coarse mesh sequences. The resulting structural correspondences are refined through semantic consistency checks and failure-case filtering. The final dataset combines accurately aligned synthetic samples with large-scale real animal videos, providing a foundation for mesh-guided animal appearance synthesis.

To reconcile structural and appearance alignment requirements, MoZoo uses two designs: Role-Aware RoPE (RAR-RoPE) for positional alignment and Asymmetric Decoupled Attention (ADA) for information routing. RAR-RoPE assigns shared temporal indices to target and mesh tokens and places asynchronous reference tokens in a non-overlapping temporal range. This aligns the target and mesh sequences without forcing frame-wise correspondence with the reference. ADA groups latent tokens by role and limits unnecessary cross-stream interactions. Target tokens attend to frame-aligned mesh guidance and appearance references, but noisy target tokens do not feed information back into the conditioning streams. This asymmetric routing reduces feature interference and unnecessary cross-stream interactions compared with full joint self-attention.

To evaluate MoZoo, we introduce **MoZooBench**, comprising 120 mesh videos with corresponding real animal videos and appearance references. Quantitative comparisons show improved appearance fidelity and structural consistency over the evaluated baselines. Figure 1 shows representative results across animal categories and motion patterns.

Our contributions are threefold:

- We introduce mesh-guided animal appearance synthesis, which generates realistic animal videos from animated meshes and multimodal appearance references.
- We construct **MoZoo-Data**, a 62K-clip synthetic-to-real dataset, and **MoZooBench**, a benchmark for evaluating appearance fidelity and structural consistency.
- We develop **MoZoo**, a video diffusion framework supporting text, image, and video references, with RAR-RoPE for temporal alignment and ADA for information routing.

2 Related Work

2.1 Controllable Video Generation

Diffusion models Ho et al. (2020); Dhariwal & Nichol (2021); Song et al. (2022) underpin video generation frameworks such as DynamiCrafter Xing et al. (2023) and VideoCrafter Chen et al. (2024), as well as large-scale video foundation models HaCohen et al. (2024); Yang et al. (2024); Zheng et al. (2024); Kong et al. (2024); Wan et al. (2025). Reference-based conditioning Zhang et al. (2023); Ye et al. (2023) improves controllability, but many reference-conditioned video methods still rely on a single image He et al. (2024); Yuan et al. (2025); Xue et al. (2025); Guo et al. (2026).

In-context DiT methods integrate conditions through token concatenation, positional encoding, and attention control. OmniGen Xiao et al. (2024) unifies text and image tokens, while Easy-Control Zhang et al. (2025) combines isolated condition processing with causal attention and KV caching. Qwen-Image Wu et al. (2025) uses semantic and reconstructive image encodings, and related methods Ju et al. (2025); Tan et al. (2025) support heterogeneous conditions within shared DiT sequences. Building on these designs, MoZoo retrieves appearance from a temporally unrelated reference video while following frame-aligned mesh geometry. RAR-RoPE encodes these distinct temporal relations, while ADA enables global appearance retrieval and frame-wise geometry access.

2.2 Hair Simulation and Rendering

High-fidelity hair production relies on specialized tools SideFX (2026), detailed shading models Chiang et al. (2015), and physics-based simulation Marschner et al. (2003); Moon et al. (2008); Zinke et al. (2008), often requiring extensive manual effort. Data-driven approaches Wang et al. (2023); Luo et al. (2024), learned strand priors Rosu et al. (2022); Sklyarova et al. (2023); Wu et al. (2024), and anisotropic Gaussian representations Zakharov et al. (2024); Zhou et al. (2024) improve hair modeling and reconstruction. ControlHair Lin et al. (2025) further demonstrates directable hair motion with video diffusion. MoZoo combines video priors with a coarse geometric proxy to reduce manual refinement in animal appearance synthesis, complementing rather than replacing physically based simulation.

3 Preliminaries

Wan2.1 Wan et al. (2025) combines a video VAE Kingma & Welling (2022) with a Diffusion Transformer (DiT) Peebles & Xie (2023) trained via Rectified Flow Liu et al. (2022). For a VAE-encoded video latent z_0, Gaussian noise ϵ, and $t \in [0, 1]$, the interpolated latent and target velocity are

$$z_t = (1 - t)z_0 + t\epsilon, \qquad u_t = \epsilon - z_0. \tag{1}$$

Given condition c (e.g., text or a reference image), the DiT learns the velocity field by minimizing

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_0,\epsilon,c} \left\| v_\theta(z_t, t, c) - u_t \right\|_2^2. \tag{2}$$

Inference integrates from $t = 1$ to $t = 0$, starting with $z_1 \sim \mathcal{N}(0, \mathbf{I})$, using the discrete update

$$z_{t-\Delta t} = z_t - v_\theta(z_t, t, c)\Delta t. \tag{3}$$

The VAE decodes the resulting z_0 into the output video.

4 METHOD

The proposed framework consists of a data construction process and a condition-aware video diffusion framework. We first describe the construction of MoZoo-Data, followed by RAR-RoPE for heterogeneous temporal alignment and ADA for information routing among target, mesh, mask, and reference streams.

4.1 MOZOO-DATA CONSTRUCTION

MoZoo-Data provides large-scale structural and appearance correspondence for mesh-guided animal appearance synthesis. Such data are scarce because animation assets are often unavailable publicly and real animal videos lack mesh representations. We combine Unreal Engine 5 (UE5) rendering Epic Games (2026) with real videos (Fig. 3). UE5 addresses the pairing bottleneck by reusing rigged assets across motions, scenes, and cameras, while real videos complement the limited appearance diversity of synthetic assets. Each training sample is a video triplet $(V_{\text{mesh}}, V_{\text{ref}}, V_{\text{tar}})$ comprising a structural motion condition, an appearance reference, and an RGB target, respectively. Mesh and target sequences depict the same shot and motion, while references need not be temporally aligned.

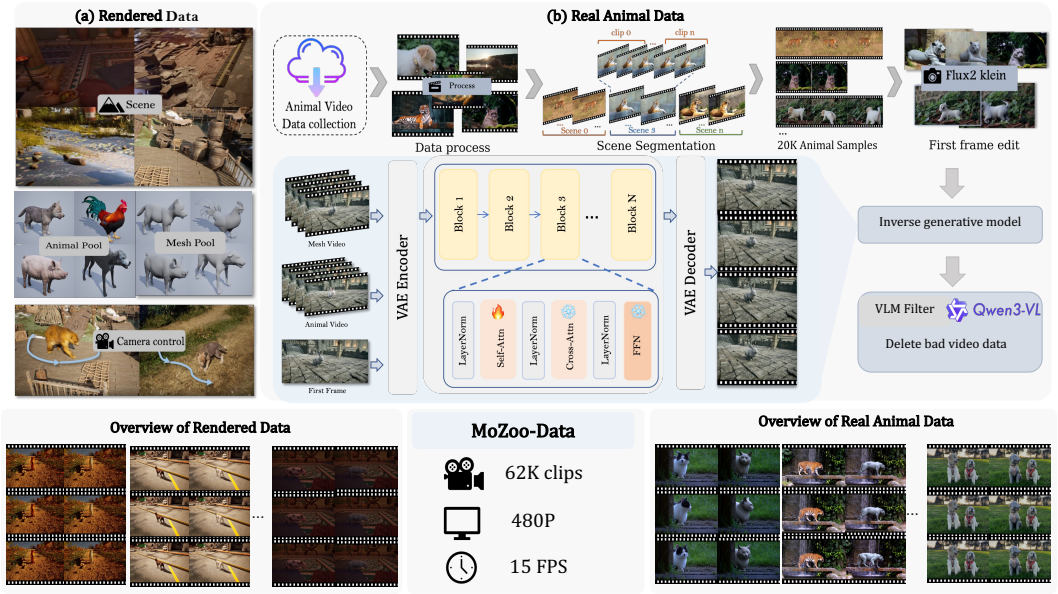

Figure 3: MoZoo-Data construction. (a) UE5 renders mesh–RGB pairs from the same animated instance. (b) Real videos are converted into mesh-conditioned pairs through first-frame appearance removal, inverse mesh estimation, instance extraction, and quality filtering. The dataset contains 62K clips at 480P and 15 FPS.

4.1.1 RENDERED DATA

We build a multi-species library of rigged, animated animal assets and an automated UE5 rendering pipeline. For each sample, we select an animal asset, an animation sequence, a 3D scene, and a camera trajectory. We first apply a uniform untextured material to render V_{mesh}. We then restore the original surface materials and bind the corresponding Groom asset to the skeletal mesh to render V_{tar}. Both passes use the same animation state and share scene and camera parameters, ensuring spatial and temporal alignment. A custom stencil pass exports per-frame animal masks M, with the first-frame mask M_{s} serving as an additional condition. We render V_{ref} from the same asset using a non-overlapping sequence with different motion. This sampling scheme preserves target appearance while preventing the model from directly copying the target motion.

4.1.2 REAL-ANIMAL DATA

Synthetic data alone cannot fully capture the appearance distribution of real animals. To reduce this domain gap, we develop a video editing pipeline converting real videos into mesh–RGB pairs. We first train a first-frame-conditioned mesh generation model on synthetic data. We collect Pexels videos Pexels (2026), standardize filenames, remove vertical videos, and resize to 1280×720. Scene detection removes shot transitions and retains semantically consistent segments without abrupt changes. We crop these segments to 81 frames. Flux2 Klein Labs (2025) removes subject appearance details from each clip's first frame to form a mesh-conditioned input. The trained mesh model then processes the edited frame and source video to produce mesh–RGB pairs. Because inverse mesh generation may fail under challenging motions, occlusions, or uncommon animal shapes, Qwen3-VL initially screens motion correspondence and temporal mesh integrity. Passing candidates undergo manual review for remaining mesh artifacts before inclusion (Appendix D). SAM 3 Carion et al. (2025) extracts instance masks and tracks the same animal across clips from one original video. We select V_{ref} from another clip of that instance showing a different action from the target, preserving appearance identity while discouraging direct motion copying. The supplementary material provides additional visualizations and representative videos.

4.2 MESH-GUIDED VIDEO DIFFUSION FRAMEWORK

Given a reference animal video and a source video featuring an untextured mesh proxy within a realistic scene, our goal is to render a lifelike subject by transferring the intricate textures from the reference video. To achieve this, we introduce **MoZoo**, a framework built upon the DiT-based video diffusion model, Wan2.1 Wan et al. (2025).

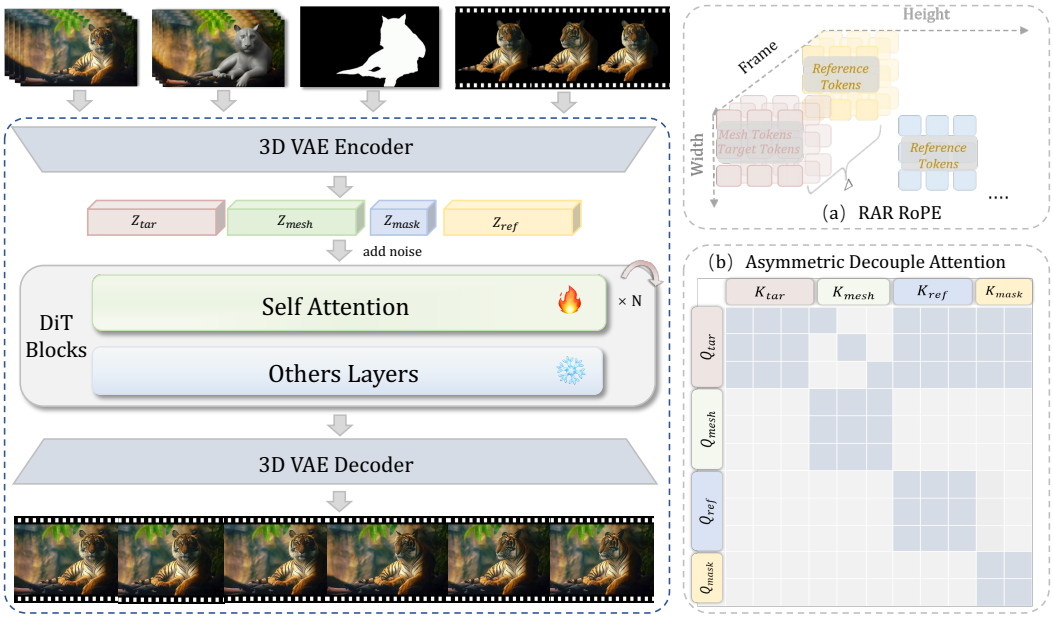

Figure 4: Overview of the condition integration framework in MoZoo. (a) RAR-RoPE assigns role-dependent temporal coordinates to heterogeneous conditions, aligning target and mesh tokens while separating asynchronous appearance references. (b) ADA establishes asymmetric information flow, allowing target queries to access guidance conditions while preventing condition streams from reading the noisy target.

4.2.1 ROLE-AWARE ROPE

Our model takes a mesh video V_mesh, an appearance reference V_ref, a first-frame subject mask M_s, and a noisy target video V_tar as visual conditions. A VAE encodes these inputs into latent tokens; flow interpolation yields the target latent during training, and Gaussian noise replaces it at inference.

After patchification, an in-context baseline concatenates visual tokens along the temporal dimension:

$$\mathbf{x} = [\mathbf{x} * \text{tgt}, \mathbf{x} * \text{mesh}, \mathbf{x} * \text{msk}, \mathbf{x} * \text{ref}], \tag{4}$$

and applies standard 3D-RoPE using global concatenation order. This imposes one temporal order on all tokens despite their different alignment requirements. The mesh and target streams require temporal synchronization, while the appearance reference only provides visual information without alignment to the target.

We propose Role-Aware RoPE (RAR-RoPE), assigning coordinates by token role rather than global position. Target and mesh tokens share temporal coordinates to preserve motion synchronization. Reference tokens are shifted by a fixed offset Δ to an independent temporal range, while the mask is anchored at temporal coordinate zero to preserve its structural role. Spatial coordinates remain unchanged to maintain the layout within each frame.

For a token i with role r, its coordinate mapping $\mathcal{M}(i, r)$ is defined as:

$$\mathcal{M}(i,r) = \begin{cases} (t_i, h_i, w_i), & \text{if } r \in \text{tgt, mesh} \\ (t_i - \Delta, h_i, w_i), & \text{if } r = \text{ref (video)} \\ (-1, h_i, w_i), & \text{if } r = \text{ref (image)} \\ (0, h_i, w_i), & \text{if } r = \text{msk} \end{cases}. \tag{5}$$

Here, (t_i, h_i, w_i) denote the local coordinates of token i after patchification. Each sample uses a single appearance modality, so image and video references are not used simultaneously.

4.2.2 Asymmetric Decoupled Attention

While RAR-RoPE encodes temporal relations, positional encoding alone does not restrict information flow. Unrestricted self-attention allows mesh and reference features to mix and condition tokens to read the noisy target, potentially interfering with the separation of structural guidance and fine-grained appearance.

We introduce **Asymmetric Decoupled Attention (ADA)** to regulate these interactions according to condition roles. Target queries aggregate heterogeneous guidance, while each condition stream remains isolated. For target frame i and condition stream $r \in \{\text{mesh}, \text{msk}, \text{ref}\}$, the accessible keys and values and the attention updates are

$$\begin{aligned} \mathcal{K}_{\text{tgt},i} &= [K_{\text{tgt}}, K_{\text{mesh},i}, K_{\text{msk}}, K_{\text{ref}}], \\ \mathcal{V}_{\text{tgt},i} &= [V_{\text{tgt}}, V_{\text{mesh},i}, V_{\text{msk}}, V_{\text{ref}}], \\ \mathbf{x}'_{\text{tgt},i} &= \text{Attn}(Q_{\text{tgt},i}, \mathcal{K}_{\text{tgt},i}, \mathcal{V}_{\text{tgt},i}), \\ \mathbf{x}'_r &= \text{Attn}(Q_r, K_r, V_r). \end{aligned} \tag{6}$$

Here, $K_{\text{mesh},i}$ and $V_{\text{mesh},i}$ contain only tokens from mesh frame i, enforcing frame-wise geometric correspondence. In contrast, reference tokens remain globally accessible, allowing appearance retrieval across reference frames without assuming temporal alignment. Target and mask tokens are also globally accessible to target queries.

Each condition stream attends exclusively to itself, preventing interactions with other conditions and feedback from the noisy target. This asymmetric pattern directs guidance toward the target while keeping condition representations decoupled. Together, RAR-RoPE and ADA provide complementary positional and attention constraints for synchronized geometry control and asynchronous appearance retrieval.

5 Experiments

5.1 Experiment Settings

Implementation details. We use Wan-2.1-T2V-14B Wan et al. (2025). We resize videos to 480×832 and sample 41 frames at 15 fps. Fine-tuning is conducted on 8 NVIDIA H200 GPUs for 20K steps, utilizing a learning rate of 2×10^{-5} and a batch size of 8.

Method	Video Quality					Paired-Target Fidelity		
	Subject Consist ↑	Background Consist ↑	Motion Smooth ↑	Imaging Qual ↑	Aesthetic Qual ↑	PSNR ↑	LPIPS ↓	SSIM ↑
VACE	93.68	96.28	99.15	69.34	55.82	15.63	0.098	0.901
VACE+Data	95.76	97.32	99.17	69.38	55.90	18.72	0.075	0.915
Refaccade	95.49	94.89	**99.24**	67.57	53.02	18.44	0.082	0.908
Channel concat. (Image)	95.54	97.69	99.13	64.21	53.02	17.89	0.090	0.893
Channel concat. (Video)	95.67	97.92	99.19	64.37	53.43	18.27	0.087	0.905
w/o RAR	97.08	98.33	99.12	69.25	55.90	18.44	0.077	0.913
w/o ADA	96.35	97.91	99.14	69.37	55.67	18.67	0.079	0.913
MoZoo (Image)	97.52	98.52	99.16	69.39	55.95	19.61	0.075	0.915
MoZoo (Video)	**97.84**	**98.72**	99.20	**69.41**	**56.30**	**20.75**	**0.070**	**0.922**

Table 2: Reference-guided results on MoZooBench. VACE, Refaccade, and MoZoo (Image) use the same reference image; MoZoo (Video) uses the proposed $\Delta = +T$ temporal offset. "+Data" and channel-concatenation baselines are trained on MoZoo-Data. Paired-target metrics are evaluated only where synchronized RGB targets are available. Best results in each column are in **bold**.

Benchmark and metrics.. To support quantitative evaluation, we introduce MoZooBench, constructed with the same pipeline as the real-world subset of MoZoo-Data using non-overlapping source videos. We compare MoZoo with VACE Jiang et al. (2025) and Refaccade Huang et al. (2025). Because both public interfaces accept only image references, we provide these methods and MoZoo (Image) with the same reference image. We conduct controlled comparisons to distinguish the contributions of training data and framework design. We fine-tune VACE on MoZoo-Data (VACE+Data) to assess the dataset contribution. We compare MoZoo with VACE+Data and matched channel-concatenation baselines to evaluate the overall framework. Ablation study isolate the contributions of RAR-RoPE and ADA. We report subject and background consistency, motion smoothness, imaging quality, and aesthetic quality for video quality Huang et al. (2024); Ke et al. (2021); Schuhmann et al. (2022). Where synchronized RGB targets are available, we measure PSNR and SSIM Wang et al. (2004) and LPIPS Zhang et al. (2018) against the paired targets. For text-guided synthesis, VQAScore Lin et al. (2024) measures prompt–output alignment.

5.2 Quantitative Comparison

Text-guided synthesis. Table 1 reports key text-guided metrics; complete results appear in Appendix Table S2. MoZoo achieves the best prompt adherence while outperforming all baselines in subject consistency, background consistency, motion smoothness, imaging quality, and aesthetic quality. These results demonstrate effective text-based appearance control without compromising structural consistency, temporal coherence, or visual quality.

Method	Subject ↑	Motion ↑	VQA ↑
VACE	93.14	99.12	0.346
VACE+Data	95.53	99.17	0.426
Channel concat.	94.49	99.04	0.457
MoZoo	**97.73**	**99.30**	**0.508**

Table 1: Text-guided results.

Reference-guided synthesis. We compare MoZoo with state-of-the-art methods using standard quantitative metrics. For a fair comparison, VACE Jiang et al. (2025) is conditioned on the input video, a reference image, and a binary mask specifying editable and fixed regions. As shown in Tab. 2, MoZoo consistently improves both structural fidelity and reference alignment. With either image (I2V) or video (V2V) references, MoZoo better matches the target appearance while preserving the structure specified by the input. The lower LPIPS further indicates closer perceptual agreement with the paired target videos. Fine-tuning VACE on MoZoo-Data substantially improves paired-target fidelity, demonstrating the value of the proposed dataset in narrowing the synthetic-to-real domain gap. Under matched training data and reference modalities, MoZoo still outperforms the channel-concatenation baseline, confirming that the improvements arise from both the data and the model design.

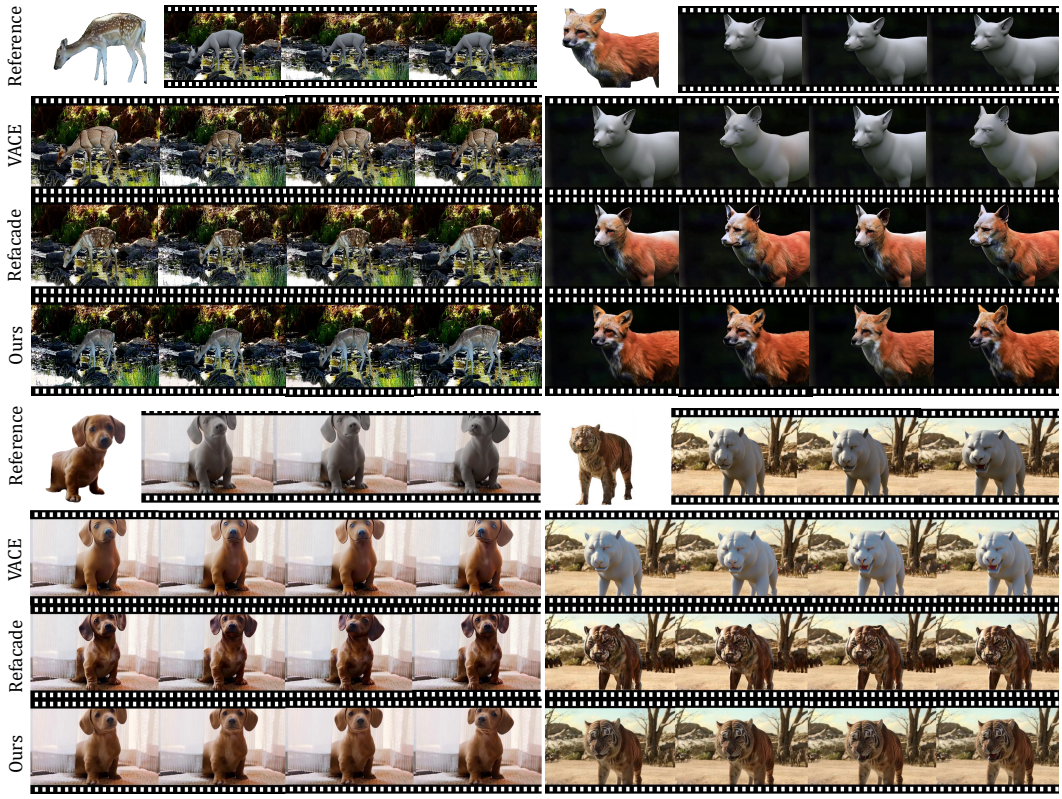

Figure 5: Qualitative comparison of reference-based generation. The results demonstrate that our proposed method can render animal subjects with superior spatio-temporal consistency, finer textural detail, and more realistic lighting effects.

5.3 Qualitative Comparison

For text-based generation, VACE can transform animal subjects in mesh-guided videos into specific species, yet its results often appear flat when representing animals with fine fur. Our method improves textural depth and realism by capturing more intricate details. Similarly, in reference-based generation, while VACE and Refacade can simulate fur to some degree, both have clear limitations. For instance, VACE occasionally fails to synthesize a coherent subject, while Refacade often struggles to map textures accurately onto the correct anatomical regions. In contrast, our approach preserves fine fur details with higher visual fidelity. Figure 5 compares reference-guided synthesis, text-guided examples are shown in Appendix Fig. S2.

5.4 Ablation Study

We evaluate our architectural components and reference modalities. Unless otherwise specified, architectural ablations use video conditioning. Table 2 reports the quantitative component ablations; Fig. 6 shows qualitative examples.

Architectural components. Removing RAR-RoPE or ADA produces distinct failure modes. Without RAR-RoPE, mesh and target tokens lose explicit positional alignment across corresponding frames, causing anatomical drift and misplaced appearance details. Removing ADA allows noisy target features to propagate into the conditioning streams, increasing interference between structural and appearance information. Consequently, fine fur textures are weakened and the generated results become blurrier.

Reference modalities. We further compare conditioning on a single image with conditioning on a reference video. Image conditioning performs comparably when the target pose is close to the ob-

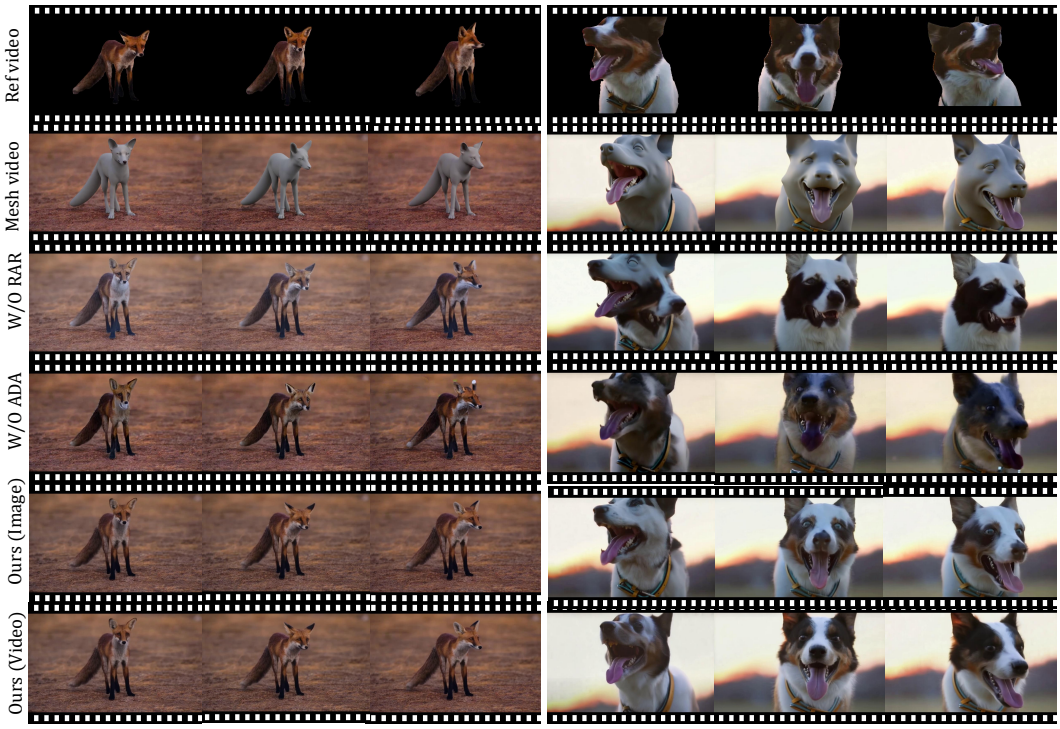

Figure 6: Ablation of temporal coordinates, condition routing, and reference modality. Removing RAR-RoPE or ADA leads to pattern misalignment or loss of fine detail in the shown examples; video reference evidence is more robust under viewpoint change.

served reference view. Under large viewpoint changes, however, a single image provides insufficient information about unseen regions, leading to missing textures and loss of detail. Video conditioning supplies appearance observations across multiple frames and viewpoints, improving identity preservation and temporal stability over a wider range of motion. Additional studies on the temporal offset δ, alternative attention designs, and the inclusion of real animal training data are provided in the supplementary material.

5.5 User Study

We recruited 25 volunteers to evaluate 20 test samples. For each criterion, they selected the preferred output based on appearance similarity to the reference, motion agreement with the input mesh, or overall quality. Table 3 reports the preference rates.

Method	Appearance	Motion	Overall
VACE	5.6%	15.0%	2.4%
Refaccade	12.2%	18.8%	10.2%
MoZoo (Ours)	**82.2%**	**66.2%**	**87.4%**

Table 3: User-study preference rates (%).

6 Conclusion

MoZoo is an in-context video diffusion framework synthesizing photorealistic animal videos from coarse mesh sequences under text, image, or video guidance. Reusing animated meshes across appearances reduces repeated grooming and material refinement. **RAR-RoPE** aligns mesh and target tokens while separating asynchronous references in temporal position space; **ADA** prevents noisy target features from contaminating conditioning streams and reduces attention cost. **MoZooBench** evaluations show strong appearance fidelity and spatiotemporal consistency across diverse animals, geometries, and motions. Complex multi-animal interactions and severe occlusions remain challenging. Future work will extend appearance synthesis to human hair and synthetic fibers with finer control over material attributes across character types. These extensions may facilitate unauthorized digital avatars or deepfake videos, risking misinformation and erosion of digital trust.

AI USE STATEMENT

Generative AI tools assisted with wording and language editing for selected passages of this manuscript. The authors take responsibility for the technical accuracy, claims, and final content of the paper.

ETHICS STATEMENT

MoZoo-Data combines animal videos rendered from UE5 assets with real animal videos collected from online sources. Our evaluation also includes a preference study with 25 volunteers; its results are reported as aggregate preference rates.

REPRODUCIBILITY STATEMENT

We describe the construction and filtering of the 62K-clip MoZoo-Data, the RAR-RoPE and ADA designs, and evaluation on the 120-case MoZooBench. For inverse mesh-video generation, we fine-tune the self-attention layers of Wan2.1 I2V-14B on UE5-rendered pairs using 81-frame 480P clips, a batch size of 8, a learning rate of 10^{-5}, and 20K steps on eight H200 GPUs. We report baseline comparisons and component ablations.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

Matt Jen-Yuan Chiang, Benedikt Bitterli, Chuck Tappan, and Brent Burley. A practical and controllable hair and fur model for production path tracing. In *ACM SIGGRAPH 2015 Talks*, SIGGRAPH '15, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336369. doi: 10.1145/2775280.2792559. URL https://doi.org/10.1145/2775280.2792559.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Epic Games. Unreal engine 5.5, 2026. URL https://www.unrealengine.com. Version 5.7.

Xu Guo, Fulong Ye, Xinghui Li, Pengqi Tu, Pengze Zhang, Qichao Sun, Songtao Zhao, Xiangwang Hou, and Qian He. Dreamid-v:bridging the image-to-video gap for high-fidelity face swapping via diffusion transformer, 2026. URL https://arxiv.org/abs/2601.01425.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation, 2024. URL https://arxiv.org/abs/2404.15275.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Youze Huang, Penghui Ruan, Bojia Zi, Xianbiao Qi, Jianan Wang, and Rong Xiao. Refaçade: Editing object with given reference texture, 2025. URL https://arxiv.org/abs/2512.04534.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention, 2025. URL https://arxiv.org/abs/2503.19907.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021. URL https://arxiv.org/abs/2108.05997.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Lingen Li, Guangzhi Wang, Zhaoyang Zhang, Yaowei Li, Xiaoyu Li, Qi Dou, Jinwei Gu, Tianfan Xue, and Ying Shan. Tooncomposer: Streamlining cartoon production with generative post-keyframing, 2025a. URL https://arxiv.org/abs/2508.10881.

Yu Li, Menghan Xia, Gongye Liu, Jianhong Bai, Xintao Wang, Conglang Zhang, Yuxuan Lin, Ruihang Chu, Pengfei Wan, and Yujiu Yang. Adaviewplanner: Adapting video diffusion models for viewpoint planning in 4d scenes, 2025b. URL https://arxiv.org/abs/2510.10670.

Weikai Lin, Haoxiang Li, and Yuhao Zhu. Controlhair: Physically-based video diffusion for controllable dynamic hair rendering, 2025. URL https://arxiv.org/abs/2509.21541.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. URL https://arxiv.org/abs/2404.01291.

Qingyang Liu, Bingjie Gao, Weiheng Huang, Jun Zhang, Zhongqian Sun, Yang Wei, Fengrui Liu, Zelin Peng, Qianli Ma, Shuai Yang, Zhaohe Liao, Haonan Zhao, and Li Niu. Animatescene: Camera-controllable animation in any scene, 2026. URL https://arxiv.org/abs/2508.05982.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Haimin Luo, Min Ouyang, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Gaussianhair: Hair modeling and rendering with light-aware gaussians, 2024. URL https://arxiv.org/abs/2402.10483.

Stephen R. Marschner, Henrik Wann Jensen, Mike Cammarano, Steve Worley, and Pat Hanrahan. Light scattering from human hair fibers. *ACM Trans. Graph.*, 22(3):780–791, July 2003. ISSN 0730-0301. doi: 10.1145/882262.882345. URL https://doi.org/10.1145/882262.882345.

Jonathan T. Moon, Bruce Walter, and Steve Marschner. Efficient multiple scattering in hair using spherical harmonics. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781450301121. doi: 10.1145/1399504.1360630. URL https://doi.org/10.1145/1399504.1360630.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Pexels. Pexels: Free stock photos & videos. https://www.pexels.com/, 2026. Accessed: 2026-01-22.

Radu Alexandru Rosu, Shunsuke Saito, Ziyan Wang, Chenglei Wu, Sven Behnke, and Giljoo Nam. Neural strands: Learning hair geometry and appearance from multi-view images, 2022. URL https://arxiv.org/abs/2207.14067.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Yahao Shi, Yang Liu, Yanmin Wu, Xing Liu, Chen Zhao, Jie Luo, and Bin Zhou. Drive any mesh: 4d latent diffusion for mesh deformation from video, 2025. URL https://arxiv.org/abs/2506.07489.

SideFX. Houdini. https://www.sidefx.com/, 2026. Accessed: 2026-03-02.

Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. Neural haircut: Prior-guided strand-based hair reconstruction, 2023. URL https://arxiv.org/abs/2306.05872.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. URL https://arxiv.org/abs/2411.15098.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Chen Cao, Jason Saragih, Michael Zollhoefer, Jessica Hodgins, and Christoph Lassner. Neuwigs: A neural dynamic model for volumetric hair capture and animation, 2023. URL https://arxiv.org/abs/2212.00613.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Keyu Wu, Lingchen Yang, Zhiyi Kuang, Yao Feng, Xutao Han, Yuefan Shen, Hongbo Fu, Kun Zhou, and Youyi Zheng. Monohair: High-fidelity hair modeling from a monocular video, 2024. URL https://arxiv.org/abs/2403.18356.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. URL https://arxiv.org/abs/2409.11340.

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Make-your-video: Customized video generation using textual and structural guidance, 2023. URL https://arxiv.org/abs/2306.00943.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-in: A lightweight and plug-and-play identity control for video generation, 2025. URL https://arxiv.org/abs/2508.07901.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition, 2025. URL https://arxiv.org/abs/2411.17440.

Egor Zakharov, Vanessa Sklyarova, Michael Black, Giljoo Nam, Justus Thies, and Otmar Hilliges. Human hair reconstruction with strand-aligned 3d gaussians, 2024. URL https://arxiv.org/abs/2409.14778.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer, 2025. URL https://arxiv.org/abs/2503.07027.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

Yuxiao Zhou, Menglei Chai, Daoye Wang, Sebastian Winberg, Erroll Wood, Kripasindhu Sarkar, Markus Gross, and Thabo Beeler. Groomcap: High-fidelity prior-free hair capture. *ACM Transactions on Graphics*, 43(6):1–15, November 2024. ISSN 1557-7368. doi: 10.1145/3687768. URL http://dx.doi.org/10.1145/3687768.

Arno Zinke, Cem Yuksel, Andreas Weber, and John Keyser. Dual scattering approximation for fast multiple scattering in hair. *ACM Trans. Graph.*, 27(3):1–10, August 2008. ISSN 0730-0301. doi: 10.1145/1360612.1360631. URL https://doi.org/10.1145/1360612.1360631.

A Additional Evaluation Experiments

Table S1 examines temporal indexing and attention visibility. For video references, placing the reference and target temporal ranges adjacent to each other ($\Delta = \pm T$) performs better than using larger or smaller offsets, consistent with reduced feature dilution. For image references, $t_{\text{img}} = -1$ gives the strongest reference alignment. Restricting q_{tgt} to k_{tgt} blocks reference injection, while omitting k_{msk} causes texture drift. Removing frame-wise mesh–target alignment increases attention computation without improving the reported metrics.

Method	Video Quality					Reference Alignment		
	Subject Consist. ↑	Backgr. Consist. ↑	Motion Smooth. ↑	Imaging Qual. ↑	Aesthetic Qual. ↑	PSNR ↑	LPIPS ↓	SSIM ↑
$\Delta = -2T$	96.87	98.32	99.19	67.70	55.53	17.42	0.076	0.915
$\Delta = -T$	97.13	98.47	**99.23**	67.61	55.67	19.14	<u>0.071</u>	<u>0.919</u>
$\Delta = -0.5T$	96.90	98.30	99.13	68.98	54.22	14.91	0.086	0.905
$\Delta = 0$	97.35	98.32	99.13	68.86	55.66	14.87	0.085	0.901
$\Delta = 0.5T$	96.70	98.27	99.13	68.28	55.25	17.03	0.075	0.915
$\Delta = 1.5T$	97.04	98.32	99.17	67.39	55.56	16.59	0.079	0.914
t_{img} = 0	97.04	97.34	**99.23**	68.89	55.01	17.89	0.082	0.911
t_{img} = T +1	97.00	98.27	99.10	69.04	<u>57.00</u>	18.64	0.082	0.910
$q_{\text{tgt}} \to k_{\text{tgt}}$	97.25	98.28	99.18	69.23	**57.01**	12.91	0.101	0.900
$q_{\text{tgt}} \to k_{\text{tgt}} + k_{\text{msk}}$	96.97	97.33	99.11	67.72	55.43	17.20	0.077	0.915
w/o frame alignment	96.90	98.27	99.17	69.01	55.04	18.34	0.0818	0.912
Ours (Img; $t_{\text{img}} = -1$**)**	<u>97.52</u>	<u>98.52</u>	99.16	<u>69.39</u>	55.95	<u>19.61</u>	0.075	0.915
Ours (Vid; $\Delta = +T$**)**	**97.84**	**98.72**	<u>99.20</u>	**69.41**	56.30	**20.75**	**0.070**	**0.922**

Table S1: Additional ablations of temporal indexing and attention visibility. Temporal-offset and attention ablations use a reference video; image-index ablations use a reference image. Bold and underlined numerical entries indicate the best and second-best distinct values among the reported rows; tied best values are both bold. Higher is better except for LPIPS.



Figure S1: Visualization of Cross-Species Migration Comparisons

A.1 Text-Guided Examples

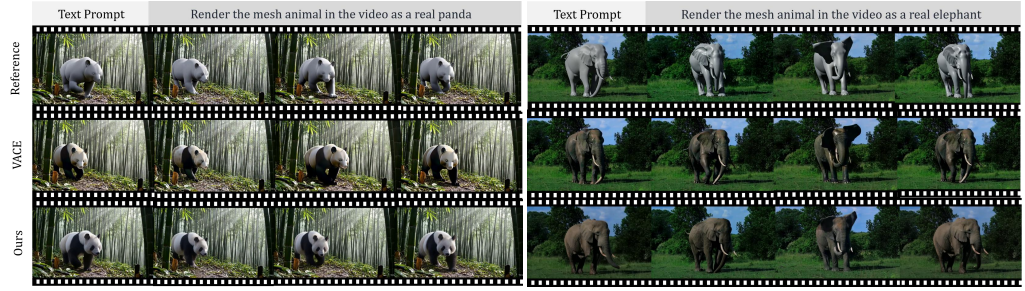

Figure S2: Text-guided synthesis with mesh motion. In these examples, MoZoo preserves fine animal textures that appear over-smoothed in VACE.

Method	Subject Consist ↑	Background Consist ↑	Motion Smooth ↑	Imaging Qual ↑	Aesthetic Qual ↑	VQAScore ↑
VACE	93.14	96.15	99.12	69.25	55.70	0.346
VACE+Data	95.53	98.12	99.17	69.39	55.80	0.426
Channel concat.	94.49	97.82	99.04	68.57	54.12	0.457
MoZoo	**97.73**	**98.32**	**99.30**	**69.47**	**56.23**	**0.508**

Table S2: Complete text-guided results. VQAScore measures prompt–output agreement. All adapted methods are trained on MoZoo-Data.

A.2 Complete Text-Guided Metrics

B More Applications

As shown in Fig. S3, our method can decouple high-level biological textures from their original anatomical structures and map them onto entirely different species. MoZoo ensures that the migrated surface attributes are biologically coherent, appearing as if they naturally originated from the target's own anatomy rather than being a superficial overlay. This zero-shot re-skinning paradigm facilitates the synthesis of high-fidelity, novel biological entities, providing an efficient and versatile tool for advanced character design and creative world-building in cinematic and interactive media.



Figure S3: Results of cross-species texture transfer. MoZoo maps biological textures from a reference animal onto a source mesh proxy of a different species, achieving photorealistic synthesis with anatomical coherence.

B.1 Animal Render

For intra-species scenarios, where the reference and target share the same biological category, our framework achieves superior performance in identity-preserving fur synthesis. As illustrated in Fig. S7, the model accurately captures and reproduces fine-grained instance-level traits, such as unique stripe distributions and subtle color gradients. By leveraging the learned spatio-temporal priors, the resulting videos demonstrate a seamless alignment between the reference textures and

the target's dynamic geometry, ensuring that the animal's distinctive visual identity is faithfully maintained throughout the sequence.



Figure S4: Visual results of cross-species fur generation. The framework maps intricate biological textures from reference videos onto target mesh proxies of disparate species, achieving high-fidelity synthesis with anatomical coherence.

B.2 Cross Species

As demonstrated in Fig. S4, our framework exhibits a robust capacity for cross-species fur generation, enabling the seamless transfer of intricate biological textures across diverse animal morphologies. Even in the presence of significant domain gaps between the reference animal and the target mesh proxy, the model faithfully preserves distinctive fur grain patterns and chromatic distributions. Beyond mere photorealism, the synthesized fur maintains strict anatomical coherence, ensuring that the projected patterns adapt naturally to the underlying muscle structures and dynamic deformations of the target.

C Dataset and Benchmark Statistics

Table S3 reports the construction and curation counts for MoZoo-Data and MoZooBench.

Subset	Quantity	Count
Rendered	Final paired clips	$\sim$30K
	Animal categories	32
	Distinct UE5 animal models	32
	Action categories	20
	UE5 scene environments	>80
	Camera trajectories	160
Real	Reconstructed candidate pairs	50K
	After Qwen3-VL screening	38K
	After manual review	32K
MoZoo-Data	Final training clips	62K
MoZooBench	Candidate cases	400
	After Qwen3-VL screening	220
	After manual review	120

Table S3: MoZoo-Data and MoZooBench curation counts. Real-data counts are reported at K precision; the approximate rendered count follows from the rounded total and real-data counts. MoZooBench case counts are reported directly.

Morphology group	Cases	Share
Quadrupedal mammals	105	87.5%
Birds	10	8.3%
Other body plans	5	4.2%
Total	120	100.0%

Table S4: MoZooBench morphology by body plan (120 cases).

MoZooBench morphology distribution. We group the 120 evaluation cases by the target animal's body plan. The quadrupedal-mammal cases include pig, cow, horse, sheep, deer, cat, dog, rabbit, fox, wolf, lion, tiger, bear, panda, zebra, giraffe, and elephant. Bird cases include peacock, duck, goose, pigeon, ostrich, and penguin. Fish, dolphin, and turtle are grouped as other body plans; turtle is separated from the mammal group despite having four limbs. The benchmark is dominated by quadrupedal mammals; the bird and other-body-plan results therefore provide limited evidence for broader anatomical generalization.

Input-mesh screening and benchmark scope. For rendered pairs, the conditioning mesh and RGB target use the same rig, animation, scene, and camera, yielding aligned pose and viewpoint by construction. Real training pairs and MoZooBench candidates both undergo the Qwen3-VL checks in Section D, followed by manual review. The automatic checks require the reconstructed mesh to follow the source animal's motion and remain complete and temporally stable throughout the clip. Human reviewers further exclude cases with fur leakage in the mesh proxy, implausible mesh deformation, or mesh instability during large or fast motions. For real training data, 38K of 50K reconstructed candidates pass Qwen3-VL screening, and manual review retains 32K. For MoZooBench, 220 of 400 candidates pass Qwen3-VL screening, and subsequent manual review retains 120 cases. The evaluation therefore measures performance on curated mesh inputs; it does not establish robustness to severely inaccurate or unstable meshes.

Training–evaluation split. We checked overlap between the retained real training clips and MoZooBench using original source-video IDs rather than cropped clip IDs. No source video is shared between the two sets.

D Mesh-Video Pair Filtering Protocol

After inverse mesh reconstruction, we apply the same Qwen3-VL screening to real training candidates and MoZooBench candidates, comparing each reconstructed mesh video with its source animal video. The filter assesses two temporal criteria. First, **motion consistency** requires the motion, pose changes, and temporal dynamics of the reconstructed mesh to remain aligned with those of the animal throughout the video. Second, **temporal mesh integrity** requires the reconstructed mesh to remain complete and structurally stable from the first frame to the last frame, without intermediate collapse, disappearance, severe flickering, abrupt deformation, or tracking failure. A candidate passes this automatic stage only when both criteria are satisfied; otherwise, it is discarded.

Among approximately 50,000 reconstructed training candidates, about 38,000 pass the automatic screen and 32,000 remain after manual review. For MoZooBench, 220 of 400 candidates pass the

same screen, and 120 remain after manual review. The manual step excludes cases with visible fur leakage into the mesh proxy, implausible mesh deformation, or mesh instability during large or fast motions.

Filtering prompt.

> You are given two temporally synchronized videos:
> Video A: the original real-animal video. Video B: the corresponding reconstructed mesh video.
> Evaluate the video pair according to the following criteria:
> 1. Motion consistency: Determine whether the reconstructed mesh follows the motion, pose changes, and temporal dynamics of the animal in Video A throughout the entire sequence.
> 2. Temporal mesh integrity: Determine whether the mesh remains complete and structurally stable from the beginning to the end of Video B. The criterion fails if the mesh collapses, disappears, flickers severely, undergoes abrupt incorrect deformation, loses tracking, or breaks at any intermediate point.
> Retain the video pair only if both criteria are satisfied. Return the result in the following format:
> { "motion_consistency": "pass" or "fail", "temporal_mesh_integrity": "pass" or "fail", "decision": "keep" or "discard", "reason": "a concise explanation" }

E Inverse Mesh-Video Generation

We train an inverse mesh-video generator on the aligned RGB and untextured mesh videos rendered by our UE pipeline. The generator uses Wan2.1-I2V-14B Wan et al. (2025) as its backbone. The first frame of a mesh video provides the image condition, while its paired RGB video provides motion and scene information. A VAE encodes the mesh first frame, RGB video, and mesh-video target. The condition tokens are concatenated with the noisy target tokens along the token sequence and processed by the DiT. We update only the self-attention layers and freeze the remaining modules.

Training follows the Wan2.1 I2V flow-matching objective. Given a clean mesh-video latent z_0, Gaussian noise ϵ, and $z_t = (1-t)z_0 + t\epsilon$, the model predicts the velocity $\epsilon - z_0$ under the concatenated visual conditions using squared-error loss. We train at 480P with 81 frames per clip, a batch size of 8, a learning rate of 10^{-5}, and 20,000 steps on eight NVIDIA H200 GPUs. Inference uses 50 sampling steps. For real animal videos, the edited mesh first frame and source RGB video condition the generation of a complete mesh-proxy video; the resulting pairs follow the filtering protocol in Section D.

F Ablation of Real Animal Data

We investigate the generalization capability of our framework when trained exclusively on synthetic data, without exposure to any real-world animal videos. As shown in Fig. S5, the model demonstrates a surprising zero-shot generalization to real-world subjects with short and sleek fur. In these cases, the latent mapping learned from synthetic samples effectively transfers to real-world surface textures, preserving the global appearance and temporal stability. However, as shown in Fig. S6, for subjects with fine, shaggy, or voluminous fur, the model encounters a significant domain gap. While it can approximate the overall color and motion, it lacks the capacity to reconstruct high-frequency, fine-grained fur strands. This indicates that while synthetic training provides a strong foundation for smooth surface alignment, the intricate, non-manifold details of complex biological fur are harder to capture without real-world supervision.



Figure S5: High-fidelity synthesis on sleek-haired real animals without real-world training supervision. The framework demonstrates robust generalization to smooth surface textures in a zero-shot setting, faithfully mapping chromatic distributions and grain patterns onto the target mesh proxy.



Figure S6: Result of real-world shaggy fur textures. Trained exclusively on synthetic data, the model captures global color and motion but struggles to recover high-frequency voluminous details, highlighting the domain gap for complex, non-manifold coat structures.



Figure S7: Results of texture transfer using real-world animal references. MoZoo faithfully maps intricate biological textures from real-world videos onto target mesh proxies, achieving high-fidelity synthesis with anatomical coherence for both sleek patterns and shaggy fur.

F.1 MULTI-ANIMAL RENDER

Multi-entity synthesis and identity mapping. Beyond single-subject transfer, our framework demonstrates robust scalability in complex multi-entity environments, supporting both "one-to-many" and "many-to-many" texture assignments. As illustrated in Fig. S8, the model can consistently broadcast the biological features of a single reference subject onto multiple target mesh proxies of different species.

In more challenging "many-to-many" scenarios where the reference video contains multiple distinct identities, our approach effectively utilizes structural and spatial cues to maintain identity fidelity. Specifically, the model performs scale-based mapping when the reference subjects exhibit significant size differences, as shown in Fig. S9, where textures are assigned based on the relative volumes of the corresponding mesh proxies. Furthermore, for interacting subjects of similar scales, the framework leverages spatial correspondence and relative positioning to ensure precise texture routing, as demonstrated by the interacting instances in Fig. S10.

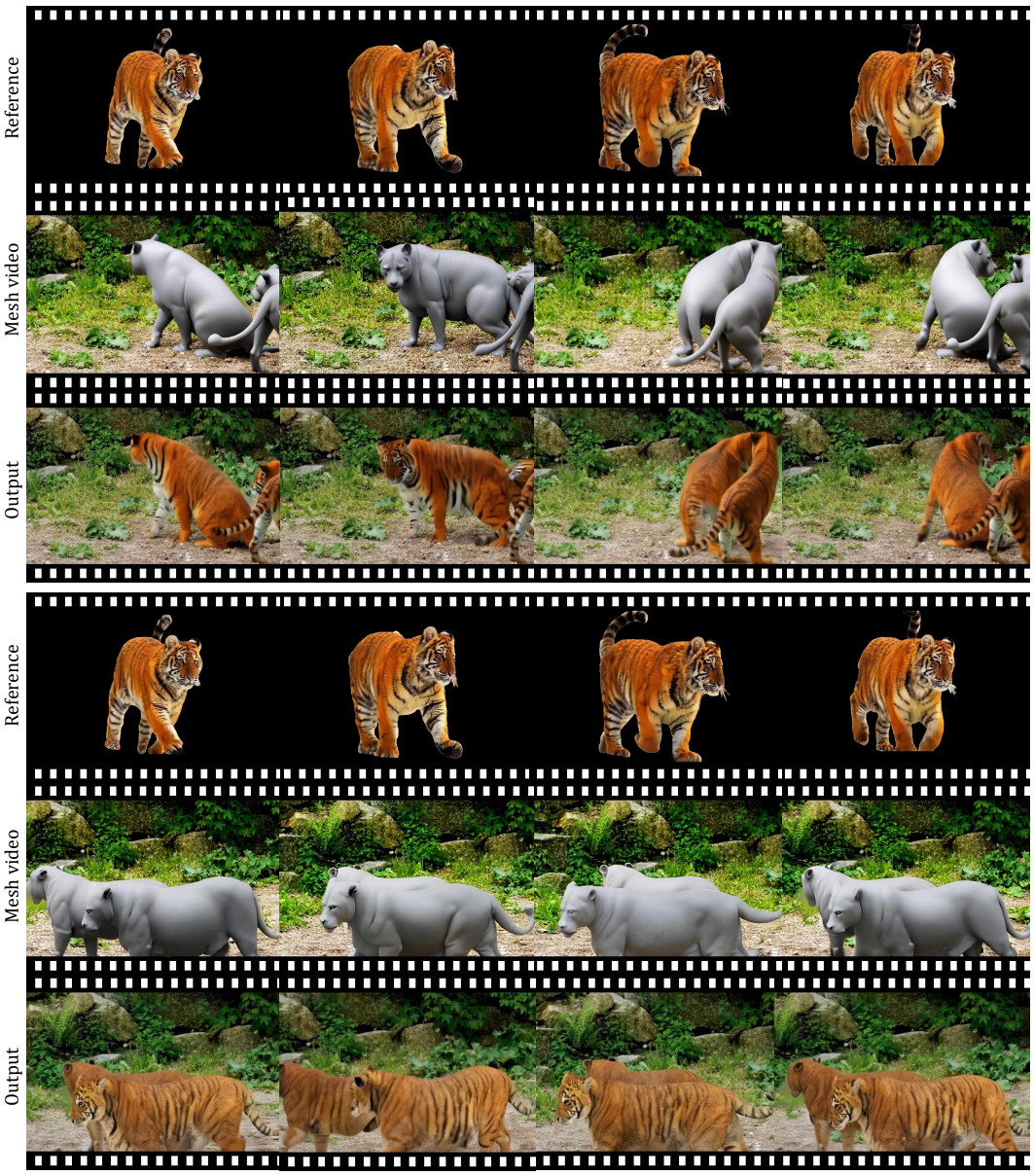

Figure S8: Results of multi-animal texture transfer. MoZoo maps biological textures from a reference animal onto multiple source mesh proxies of different species, achieving photorealistic synthesis with anatomical coherence across all subjects.

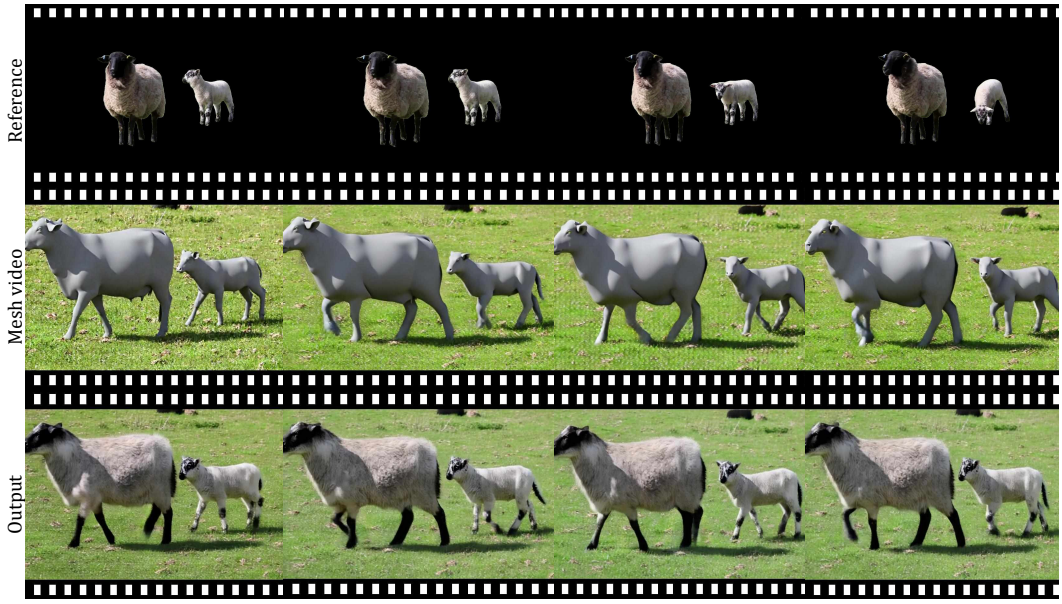

Figure S9: Multi-entity rendering with scale-based mapping. Identity consistency is preserved by correlating the relative sizes of reference subjects with their corresponding target mesh proxies.

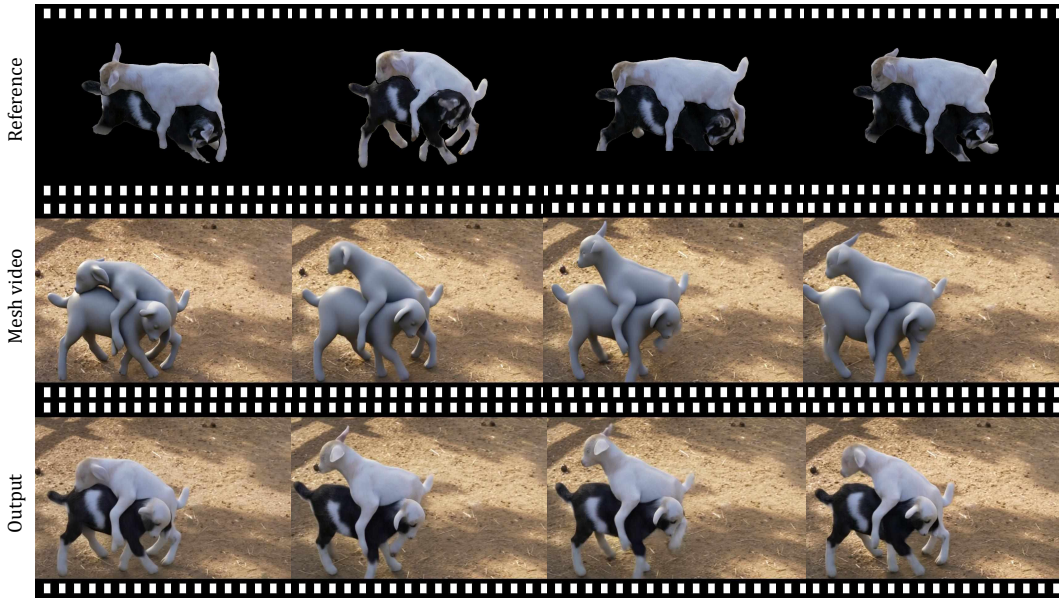

Figure S10: Multi-animal synthesis with spatial correspondence. The framework leverages relative positional cues from the reference video to accurately assign textures to interacting subjects with overlapping geometries.

G LIMITATION

The effectiveness of our multi-animal texture transfer relies heavily on the structural and spatial correspondence between the reference video and the target mesh proxies. While our model excels at one-to-many mapping—where a single reference is consistently applied to multiple instances—it also succeeds in spatially guided scenarios where the reference video provides clear cues for interacting subjects, as demonstrated in Fig. S10. However, a significant bottleneck arises when handling disparate identities that lack explicit spatial or semantic mapping to the target scene. As illustrated in Fig. S11, when provided with heterogeneous references such as a tiger and a cougar for separate target meshes, the model struggles to independently route textures to their intended instances. This often leads to identity blending, producing a blurred hybrid texture that merges features from both species rather than performing discrete instance-level assignments. Such a limitation highlights the necessity for sophisticated spatially aware routing mechanisms to achieve independent texture control in complex multi-entity environments.

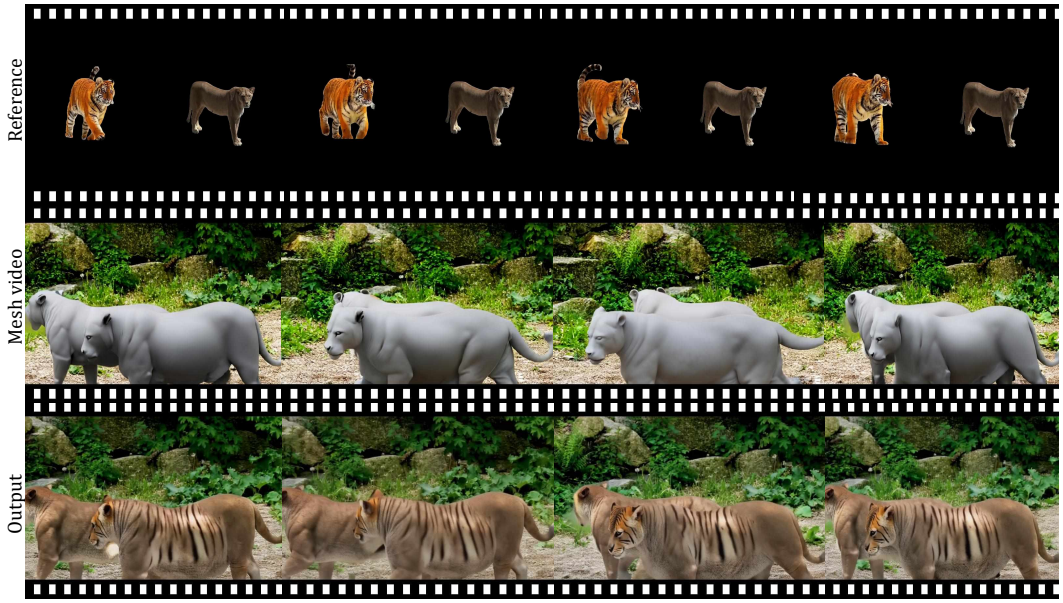

Figure S11: Failure case of multi-animal mapping due to a lack of structural correspondence. The absence of matching spatial positions and relative scales between the reference subjects and target mesh proxies leads to identity blending and incorrect texture assignment.

H FUTURE WORK

Despite the promising results of MoZoo, several avenues remain for future exploration. First, we plan to transition from single-subject scenarios to complex multi-object environments, exploring how to maintain spatio-temporal consistency and individual identity across multiple interacting entities. Second, we intend to investigate the simulation of intricate biological structures, such as the iridescent feathers of a peacock. Capturing these fine-grained details requires more sophisticated physics-aware priors to model structural coloration and complex anisotropic reflections, which remains a compelling frontier for our work.