Text–Image Conditioned 3D Generation

Supplementary Material

A. Contents

In this supplement, we provide the following contents:

- The design of the cross-modal late-fusion module used for ablation.
- The dataset composition of TRELLIS-500K.
- The evaluation metric definitions and rendering settings.
- Compatibility between TIGON and TRELLIS.
- More qualitative results.
- Full text prompts used in the main paper.

B. Design of Ablation for the Cross-Modal Late-Fusion Module

To assess whether more sophisticated mechanisms can outperform our simple averaging strategy, we design two learnable late-fusion variants.

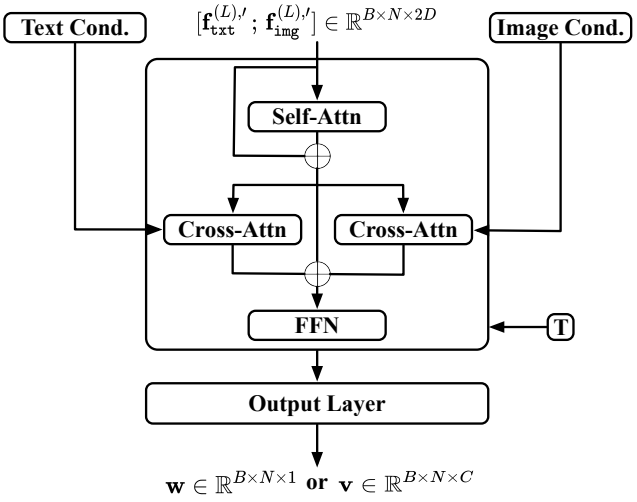

Figure A1. The adaptive fusion module used in the ablation study. D denotes the intermediate feature dimension, and C denotes the latent output dimension.

A fusion module must access the current branch predictions $\mathbf{f}_{\texttt{txt}}^{(L)}$ and $\mathbf{f}_{\texttt{img}}^{(L)}$, the modality conditions $\mathbf{T}$ and $\mathbf{I}$, and the denoising timestep t. To integrate these inputs, we employ a dual cross-attention module. As shown in Fig. A1, we concatenate the two branch features along the channel dimension and obtain a fused representation:

$$\mathbf{f}_{\texttt{fused}} = \mathcal{M}\big([\mathbf{f}_{\texttt{txt}}\,;\,\mathbf{f}_{\texttt{img}}],\, t,\, \mathbf{T},\, \mathbf{I}\big). \tag{1}$$

Different output heads applied to $\mathbf{f}_{\texttt{fused}}$ produce the two fusion variants below.

Linear Fusion with Learnable Adaptive Weighting (AW). To produce element-wise fusion weights, the fusion module applies a linear projection $\mathbb{R}^{B\times N\times 2D} \to \mathbb{R}^{B\times N\times 1}$, followed by a sigmoid activation to obtain $\mathbf{w} \in \mathbb{R}^{B\times N\times 1}$. The final fused prediction is computed as

$$\mathbf{v} = \mathbf{w}\,\mathbf{v}_{\texttt{txt}} + (1-\mathbf{w})\,\mathbf{v}_{\texttt{img}}. \tag{2}$$

Attention-Based Cross-Modal Fusion (AT). Instead of predicting fusion weights, this variant lets the fusion module directly generate modality-specific features. To maintain training stability, we reuse the original output projections $\mathcal{G}_* : \mathbb{R}^D \to \mathbb{R}^C$ for $* \in \{\texttt{txt}, \texttt{img}\}$. We split the fused representation along the channel dimension to obtain $\hat{\mathbf{f}}_{\texttt{txt}}, \hat{\mathbf{f}}_{\texttt{img}} \in \mathbb{R}^{B\times N\times D}$. Each is then projected to the latent dimension, and the final prediction is formed by summation:

$$\mathbf{v} = \frac{1}{2}(\mathcal{G}_{\texttt{txt}}(\hat{\mathbf{f}}_{\texttt{txt}}) + \mathcal{G}_{\texttt{img}}(\hat{\mathbf{f}}_{\texttt{img}})). \tag{3}$$

Note that we omit the normalization operation for brevity.

C. TRELLIS-500K Dataset Overview

TRELLIS-500K is a large-scale 3D asset collection assembled in prior work [9], drawing from several publicly available repositories. The dataset merges objects from Objaverse-style sources together with high-quality CAD and artist-designed assets, followed by filtering procedures to remove models with missing geometry or severely degraded textures. Each object is additionally paired with a detailed natural-language caption generated using GPT-4o [6], providing consistent semantic supervision for text-driven 3D generation.

Objaverse-Derived [2, 3]. A substantial portion of TRELLIS-500K comes from higher-quality subsets of Objaverse-XL, particularly assets originating from Sketchfab (Objaverse V1) and selected GitHub contributions. These models cover a wide range of manually designed shapes, photogrammetry scans, and professionally captured artifacts. Lower-quality objects from the broader Objaverse-XL collection are excluded.

ABO [1]. ABO contributes a set of professionally authored household product models characterized by clean topology and high-resolution materials, enriching the dataset with well-designed, manufacturable assets.

3D-FUTURE [4]. 3D-FUTURE provides industrial-grade furniture models with detailed geometry and realistic textures, complementing other sources with contemporary interior designs.

HSSD [7]. Assets from HSSD include indoor objects such as decorative items and furnishings originally curated for embodied AI research. These assets are structurally consistent and help broaden the dataset's coverage of indoor categories.

Overall, TRELLIS-500K offers a curated mixture of diverse, reasonably clean 3D assets with high-quality textual descriptions, and serves as a strong large-scale dataset for training text- and image-conditioned 3D generative models.

D. Rendering Settings and Evaluation Metric Definitions

D.1. Rendering Settings

To assess generation quality, we render four reference views for each ground-truth object using cameras placed at yaw angles of $0°, 90°, 180°, 270°$ and a fixed pitch of $30°$, all looking toward the origin with a $40°$ field of view and positioned uniformly on a sphere of radius 2. We apply the same rendering protocol to the corresponding generated object to obtain its synthesized views. Image features are then extracted using the CLIP image encoder and DINOv2 to compute the CLIP similarity score and $\text{FD}_{\texttt{DINOv2}}$, respectively.

D.2. Evaluation Metric Definitions

CLIP. For each instance, we render four views for the ground-truth (GT) object and four views for the generated object, and compute the cosine similarity between every GT–generated pair, yielding a 4×4 similarity matrix. Since the generated objects are not canonicalized with respect to front/back orientation, we do not know the exact correspondence between GT and generated views. Therefore, we apply the Hungarian matching to this 4×4 matrix to find the optimal one-to-one assignment, and use the resulting matching score (average cosine similarity over the matched pairs) as the final CLIP score.

Fréchet Inception Distance (FID). To assess distributional similarity between real and generated objects, we compute a Fréchet Distance in the DINOv2 feature space. All GT and generated renders are embedded using a pre-trained DINOv2 encoder to obtain sets of "real" features $\{\mathbf{x}_i\}$ and "generated" features $\{\mathbf{y}_i\}$. We estimate the empirical means and covariances $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ and $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ of the two distributions, and report

$$\text{FD}_{\texttt{DINOv2}} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \text{Tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}\Big), \tag{4}$$

where lower values indicate that generated features more closely match the distribution of GT features.

ULIP and Uni3D. To evaluate the geometric fidelity of the generated 3D object, we employ ULIP [10] and Uni3D [11] to measure semantic consistency between the generated 3D shape and the corresponding GT renderings. For each instance, we encode the four reference views using the ULIP or Uni3D image encoder, and encode the generated mesh using the corresponding point-cloud encoder applied to its vertices. The average cosine similarity between the image embeddings and the point-cloud embedding is reported as the ULIP or Uni3D score.

E. Compatibility between TIGON and TRELLIS

In this paper, we instantiate TIGON on UniLat3D, a single-stage extension of TRELLIS. However, the core idea of TIGON is not restricted to UniLat3D. To demonstrate this, we conduct an additional experiment directly on TRELLIS: we integrate the TRELLIS text and image models within the TIGON framework and evaluate the resulting model.

Due to resource limitations, we add cross-modal bridges only to the sparse-structure flow of TRELLIS and fine-tune it, while keeping the SLAT flow as a simple fusion of the two modality branches.

The performance is reported in Table A1. With cross-modal fusion, TRELLIS exhibits improved generation quality compared with its single-modality variants. These results highlight the potential of TIGON as a general multi-modal fusion framework that is compatible with a broader class of flow-based 3D generators.

Model	Cond.	CLIP	$\text{FD}_{\texttt{DINOv2}}$
TRELLIS	I	90.50	98.75
TRELLIS	T	86.30	148.21
TRELLIS (w/o ss-bridge)	I+T	91.23	80.35
TRELLIS (w/ ss-bridge)	I+T	91.51	75.35

Table A1. Integrating TIGON with TRELLIS. Experiment is conducted on Toys4K. We use 3D-GS as the representation. 'ss-bridge' denotes the cross-modal bridge for the sparse-structure flow model.

F. Additional Qualitative Results

Mesh Generation. In the main paper, we provide visualization with 3D-GS representation. We further pro-

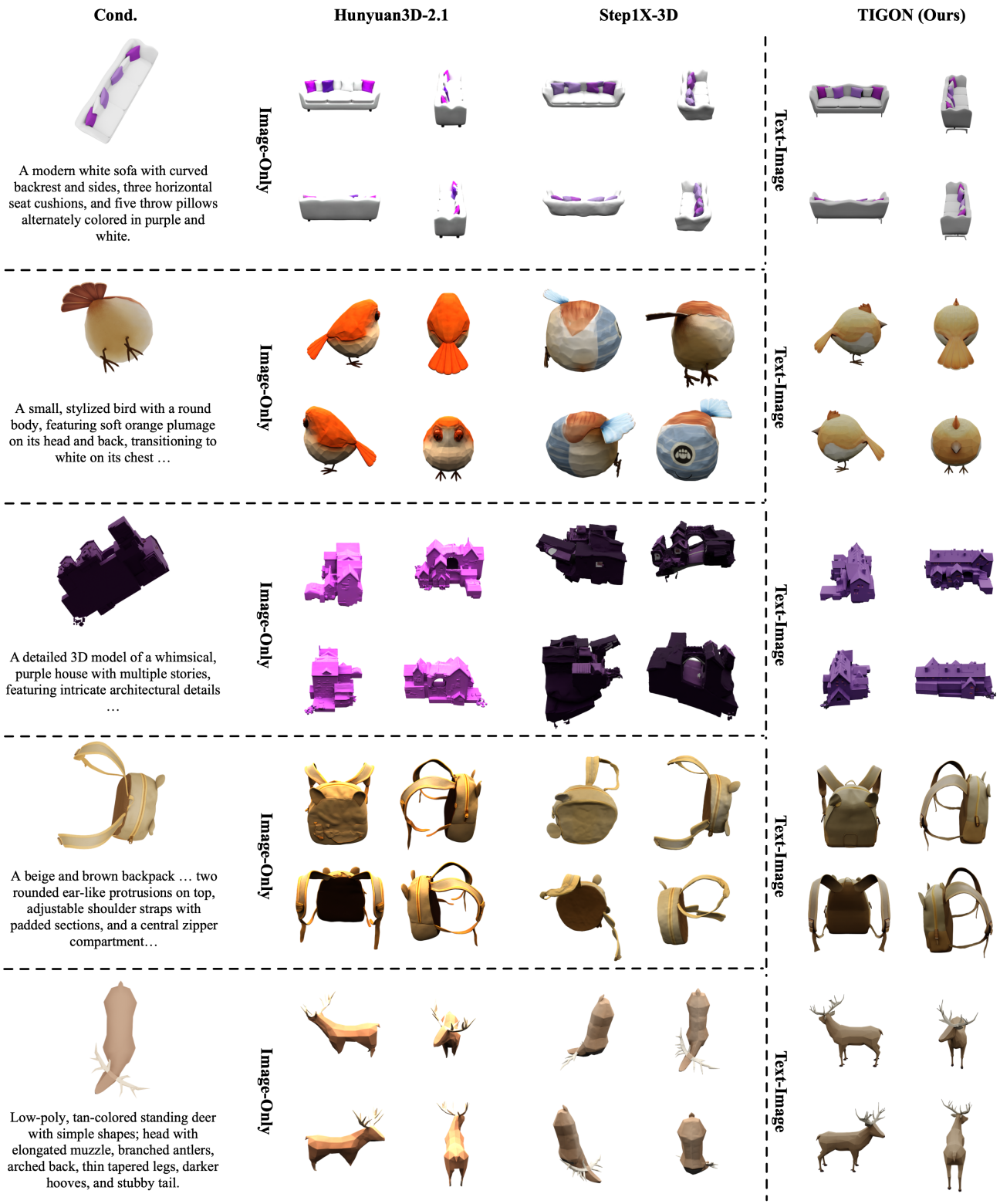

Figure A2. Meshes generated by Hunyuan3D-2.1, Step1X-3D, and TIGON. For the full text prompts, please refer to Sec. G.

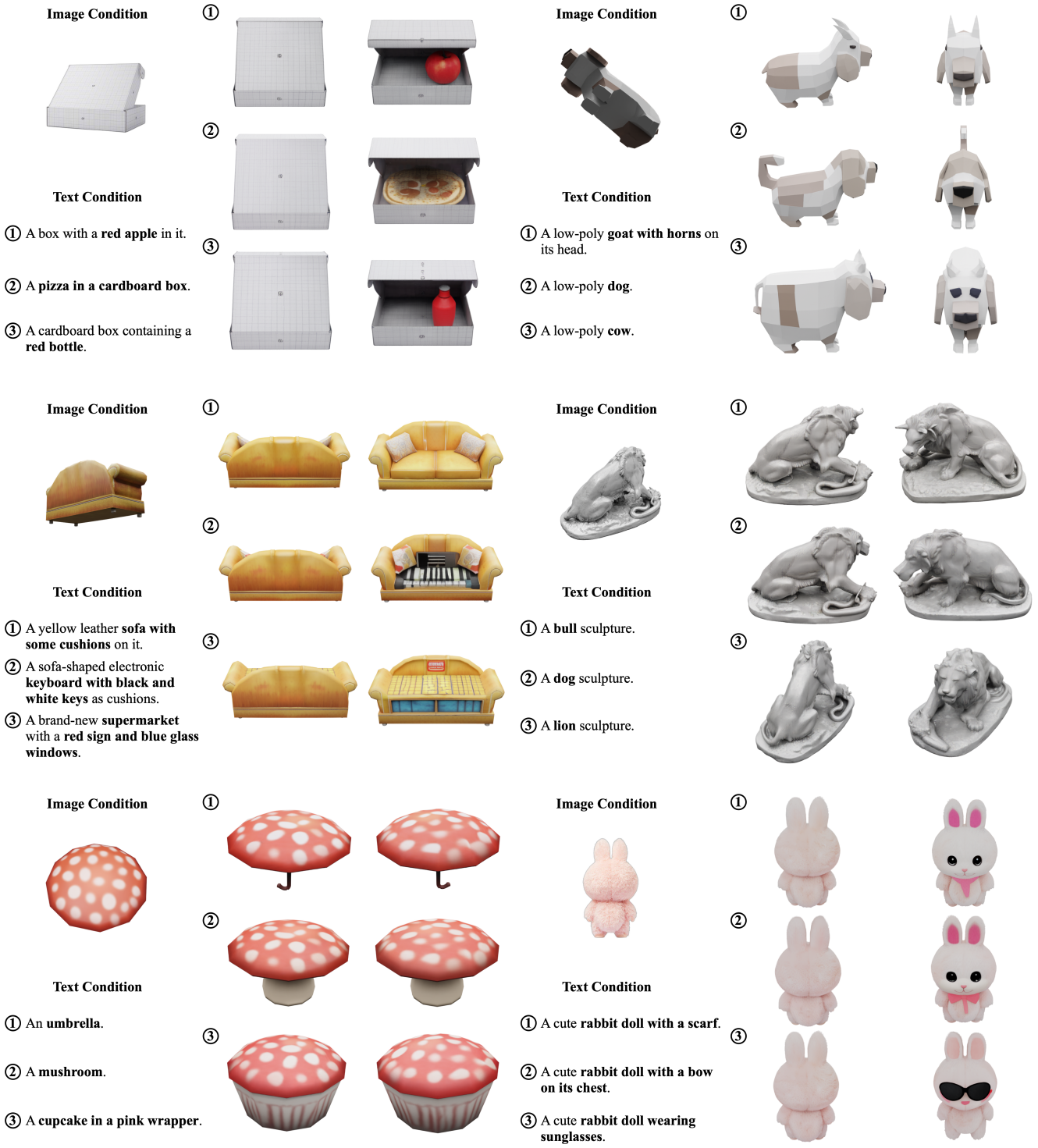

Figure A3. Additional visualization results of controllable generation with TIGON under joint text and image conditioning.

vide comparison results with other mesh generation methods [5, 8] to demonstrate the ability of mesh generation of TIGON. Existing approaches exhibit a strong dependence on favorable viewpoints. For example, in the second row of Fig. A2, although the input image clearly depicts a bird, Step1X-3D fails to produce a plausible geometry. Similarly, in the fifth row, Hunyuan3D-2.1 cannot generate the deer's legs. With explicit semantic guidance from

text, TIGON successfully reconstructs these cases, producing meshes that better align with both appearance cues and object semantics.

Controllable Generation. Additional results in Fig. A3 further illustrate TIGON's controllable generation ability. By combining different condition images with different text prompts, TIGON produces diverse 3D objects while maintaining strong visual alignment, highlighting the flexibility and expressiveness enabled by joint text–image conditioning. We also provide videos that more fully show these controllable generation results; please refer to the supplementary materials.

G. Full Text Prompts List

We provide here the full text prompts that are abbreviated in our figures.

Fig. 4: From the 1st to the 5th row:

- A whimsical, yellow toy-like car with a clown driver, blue side panels, and roof, red circles near the wheels, three roof balloons, a yellow wind-up key, light blue bumpers, and a weathered, vintage aesthetic.
- A circular dartboard with concentric rings alternating between white and red, featuring a central bullseye and a dart embedded in the outer red ring. The dart has a black shaft, a brown tip, and red fletching. The dartboard's surface appears to have a matte finish, and the lighting highlights its three-dimensional form.
- A handheld gaming console with a sleek black body, featuring two detachable controllers, one blue on the left and one red on the right. The blue controller has a directional pad and four action buttons labeled A, B, X, and Y. The red controller includes a joystick and additional buttons. The device has a glossy finish with visible screws and ports along its edges, indicating a portable design for gaming on the go.
- A golden trophy with intricate engravings and detailed textures, featuring two handles, a transparent lid with a purple band, and blue ribbons with logos draped over it. The base has multiple tiers with inscriptions, and the overall design includes reflective surfaces and polished finishes.
- A vintage toaster with a compact, rectangular shape, rounded edges, predominantly orange with metallic silver accents, two top bread slots, front and back panels featuring curved orange with central silver panels, inwardly curved side panels with a metallic lever and two knobs, and a wider base.

Fig. 6: A modern all-in-one desktop computer with a sleek, white stand, slightly curved white back, flat rectangular screen with rounded corners, thin black side and top bezels, and a thicker white bottom bezel.

Fig. A2: From the 1st to the 5th row:

- A modern white sofa with curved backrest and sides, three horizontal seat cushions, and five throw pillows alternately colored in purple and white.
- A small, stylized bird with a round body, featuring soft orange plumage on its head and back, transitioning to white on its chest. The bird has a short, pointed black beak, dark eyes, and delicate brown legs with clawed feet. Its tail is short and slightly fanned, matching the orange coloration of its back. The texture appears smooth and slightly fluffy, giving it a plush, cartoon-like appearance.
- A detailed 3D model of a whimsical, purple house with multiple stories, featuring intricate architectural details such as gabled roofs, chimneys, windows with varying sizes and shapes, and a staircase leading to the entrance. The house has a textured surface resembling stone or brick, with small decorative elements like railings and a balcony. The surrounding area includes scattered debris and a small figure near the base, adding context to the scene.
- A beige and brown backpack with a textured fabric surface, featuring two rounded ear-like protrusions on top, adjustable shoulder straps with padded sections, and a central zipper compartment. The backpack has a structured design with visible stitching details and reinforced areas around the straps and zippers.
- Low-poly, tan-colored standing deer with simple shapes; head with elongated muzzle, branched antlers, arched back, thin tapered legs, darker hooves, and stubby tail.

References

[1] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, 2022. 1

[2] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 1

[3] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 2024. 1

[4] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *IJCV*, 2021. 2

[5] Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025. 4

[6] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[7] Mukul Khanna*, Yongsen Mao*, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation. *arXiv preprint arXiv:2306.11290*, 2023. 2

[8] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 4

[9] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *CVPR*, 2025. 1

[10] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *CVPR*, 2023. 2

[11] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 2