Affostruction: 3D Affordance Grounding with Generative Reconstruction

Supplementary Material

A. Implementation details

A.1. Model architectures

Our method employs two flow-based models: a flow transformer for multi-view 3D reconstruction (Stage 1) and a sparse flow transformer for affordance grounding (Stage 2). Both follow the rectified flow framework [3, 8] but operate on different representations.

Flow transformer (Stage 1). The flow transformer extends TRELLIS [8] to support multi-view RGBD inputs through sparse voxel fusion conditioning. It processes a dense noise tensor $\mathbf{X} \in \mathbb{R}^{16^3 \times 8}$ (4096 tokens) conditioned on fused DINOv2 [5] features from multiple views. We use DINOv2-ViT-L/14 with registers (dinov2_vitl14_reg) as the visual feature extractor, producing 1024-dimensional features. Complete architectural specifications are provided in Table A1.

Sparse flow transformer (Stage 2). The sparse flow transformer operates on sparse voxel positions predicted by Stage 1, generating affordance heatmaps conditioned on natural language queries. We use CLIP-ViT-L/14 [6] (openai/clip-vit-large-patch14) as the text encoder, producing 768-dimensional text embeddings. Unlike the dense reconstruction model, this sparse formulation processes only occupied voxels ($L \ll 4096$), enabling efficient affordance prediction. Full specifications are provided in Table A2.

A.2. Training configuration

Common training setup. Both models share core training parameters: 450K steps with batch size 8 per GPU, AdamW optimizer [4] (learning rate 10^{-4}, no weight decay), EMA (rate 0.9999), and mixed precision (FP16) with adaptive gradient clipping (max norm 1.0, 95th percentile). Timestep t is sampled from a logit-normal distribution ($\mu = 1.0, \sigma = 1.0$). We apply 10% unconditional training for classifier-free guidance [1].

Stage 1: Multi-view reconstruction. The reconstruction model is trained with MSE loss and validated on Toys4k [7] (primary metric: MSE). During training, we randomly sample 1–8 views per object to ensure robust multi-view fusion at inference time.

Stage 2: Affordance grounding. The affordance model is trained with a combination of binary cross-entropy and Dice loss, suited for the binary nature of affordance labels. We use elastic training with a linear memory controller (target ratio 0.75) to handle variable structure sizes. The noise scale is set to 5.0 to account for the binary distribution of affordance heatmaps. Validation uses Affogato [2] (primary metric: average IoU).

A.3. Evaluation configuration

3D reconstruction on Toys4k [7]. Following TRELLIS [8], we evaluate reconstruction quality on 1,250 randomly selected samples (SHA256 identifiers in Table A3). For image metrics, we render both ground truth and predictions from a fixed viewpoint (radius $r = 2.0$, FOV 40, resolution 512×512) to compute PSNR and LPIPS. For point cloud metrics (Chamfer Distance and F-score), we render depth maps from 100 views uniformly distributed via Hammersley sampling, unproject to 3D coordinates, and sample 100K points.

Affordance grounding on Affogato [2]. We evaluate on the entire test split, following standard protocol: first view and first query per sample. Since our method is generative, we use a reduced noise scale of 0.5 at inference (compared to 5.0 during training) to obtain more consistent predictions for quantitative evaluation.

B. Sampling parameters search

We search for the optimal number of sampling steps for the multi-view reconstruction model. Figure A1 shows volumetric IoU across different sampling steps (1, 5, 10, 15, 20) and number of input views (1–6). Reconstruction quality plateaus at 5 steps regardless of the number of views, with additional steps providing diminishing returns at increased computational cost. At 5 steps, our model achieves a sampling time of approximately 0.25 seconds, which is $5\times$ faster than the 25-step default (1.29s) of TRELLIS [8]. This efficiency is important for active view selection, where rapid reconstruction enables iterative viewpoint refinement in robotic applications.

C. Additional qualitative results

Figure A2 illustrates the iterative refinement process of Affostruction starting from challenging initial observations. We select starting viewpoints where target functional areas have minimal visibility to test the system under difficult conditions. As views are actively selected based on predicted affordances, the accumulated observations lead to more complete reconstruction, which in turn enables more accurate affordance prediction on the recovered geometry. Both geometric quality and affordance localization progressively improve as more informative views are incorporated.

Table A1. **Flow transformer architecture (Stage 1: Multi-view Reconstruction).** The model generates dense 3D structure from multi-view RGBD observations through DINOv2 sparse voxel fusion conditioning.

Component	Value	Description
Transformer Architecture		
Resolution	16	Spatial resolution of dense 3D grid ($16^3 = 4096$ tokens)
Input channels	8	Channels of input noise tensor
Output channels	8	Channels of denoised output tensor
Model channels	768	Hidden dimension of transformer blocks
Conditioning channels	1024	Dimension of DINOv2 features (ViT-L/14)
Number of blocks	12	Depth of DiT (Diffusion Transformer) backbone
Number of heads	12	Multi-head attention heads per block
MLP ratio	4	Hidden dimension multiplier for feed-forward layers
Patch size	1	Spatial patch size for tokenization
Positional encoding	APE	Absolute positional encoding
QK RMS norm	✓	RMS normalization for query-key projections
Precision	FP16	Mixed precision training with FP16
Conditioning Mechanism		
Visual encoder	DINOv2-ViT-L/14-reg	Feature extractor for RGBD views
Feature dimension	1024	Output dimension of DINOv2 features
Voxel resolution	16	Resolution for sparse voxel fusion
Image size	224×224	Input image resolution for DINOv2
Max views	8	Maximum number of views during training
Fusion method	Average	Feature averaging for overlapping voxels

Table A2. **Sparse flow transformer architecture (Stage 2: Affordance Grounding).** The model generates affordance heatmaps on sparse 3D structure conditioned on text queries via CLIP.

Component	Value	Description
Transformer Architecture		
Resolution	64	Spatial resolution for latent representation
Input channels	1	Single channel for affordance heatmap
Output channels	1	Single channel affordance prediction
Model channels	768	Hidden dimension of transformer blocks
Conditioning channels	768	Dimension of CLIP text embeddings (ViT-L/14)
Number of blocks	12	Depth of DiT backbone
Number of heads	12	Multi-head attention heads per block
MLP ratio	4	Hidden dimension multiplier for feed-forward layers
Patch size	2	Spatial patch size for tokenization
I/O residual blocks	2	Number of input/output residual blocks
I/O block channels	128	Hidden channels in I/O residual blocks
Positional encoding	APE	Absolute positional encoding
QK RMS norm	✓	RMS normalization for query-key projections
Precision	FP16	Mixed precision training with FP16
Conditioning Mechanism		
Text encoder	CLIP-ViT-L/14	Text feature extractor (openai/clip-vit-large-patch14)
Feature dimension	768	Output dimension of CLIP text embeddings

Table A3. **Toys4k test set samples (SHA256 identifiers).** The 1,250 samples used for 3D reconstruction evaluation, following TRELLIS [8]. Full list available in `toys4k_test_ids.txt`. Hashes truncated to first 12 characters for display.

SHA256 Object Identifiers (1,250 samples)				
000a283e3a4e...	002d00832905...	0036c7bf5fa3...	00b614f80a13...	0100555a135f...
016be2974e32...	019335038b79...	01a79ca24eac...	01ac5979fed3...	02065ccd7123...
021c0a67be93...	0262655e3219...	0268f36995da...	0289dd8d108d...	0290334c3684...
02a87d37f648...	02ba6532f9de...	02c70213d5af...	02e84388b24c...	02e9faa6bff3...
...(1,230 additional samples)				

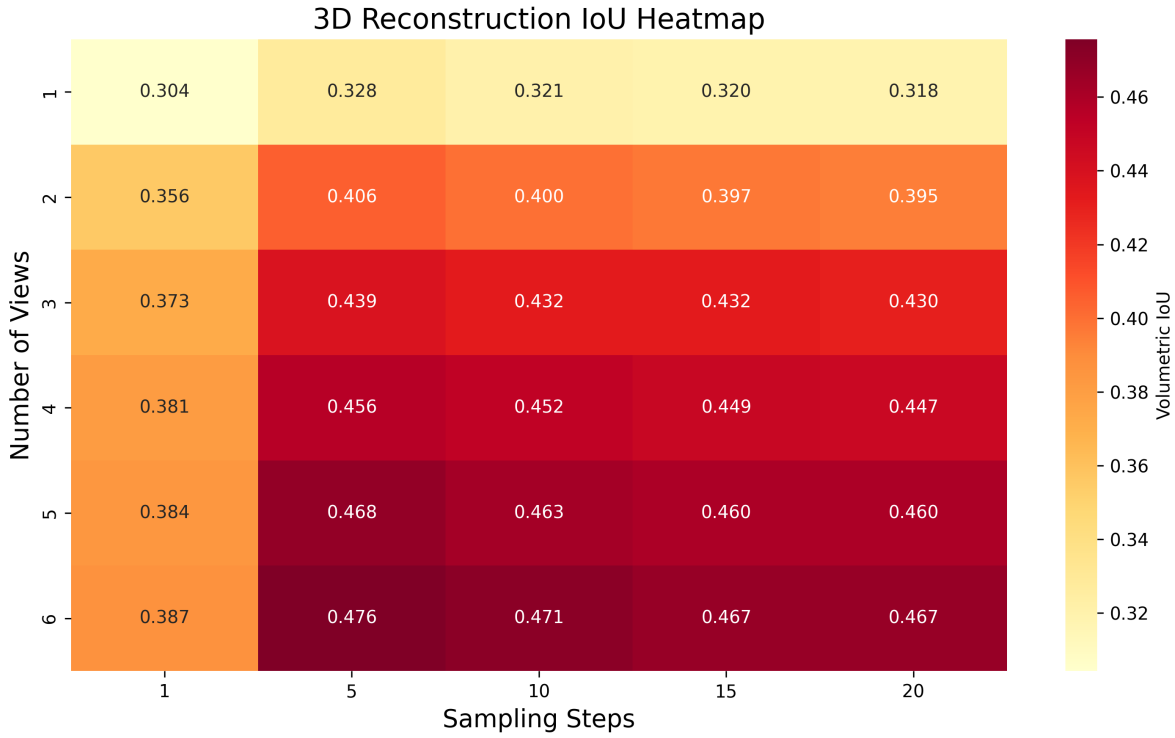

Figure A1. **Sampling step ablation across different number of views.** We evaluate volumetric IoU for varying sampling steps (1, 5, 10, 15, 20) with 1–6 input views. Reconstruction quality saturates at 5 steps across all view configurations, achieving 5× faster sampling (0.25s) compared to the default 25 steps in TRELLIS [8].

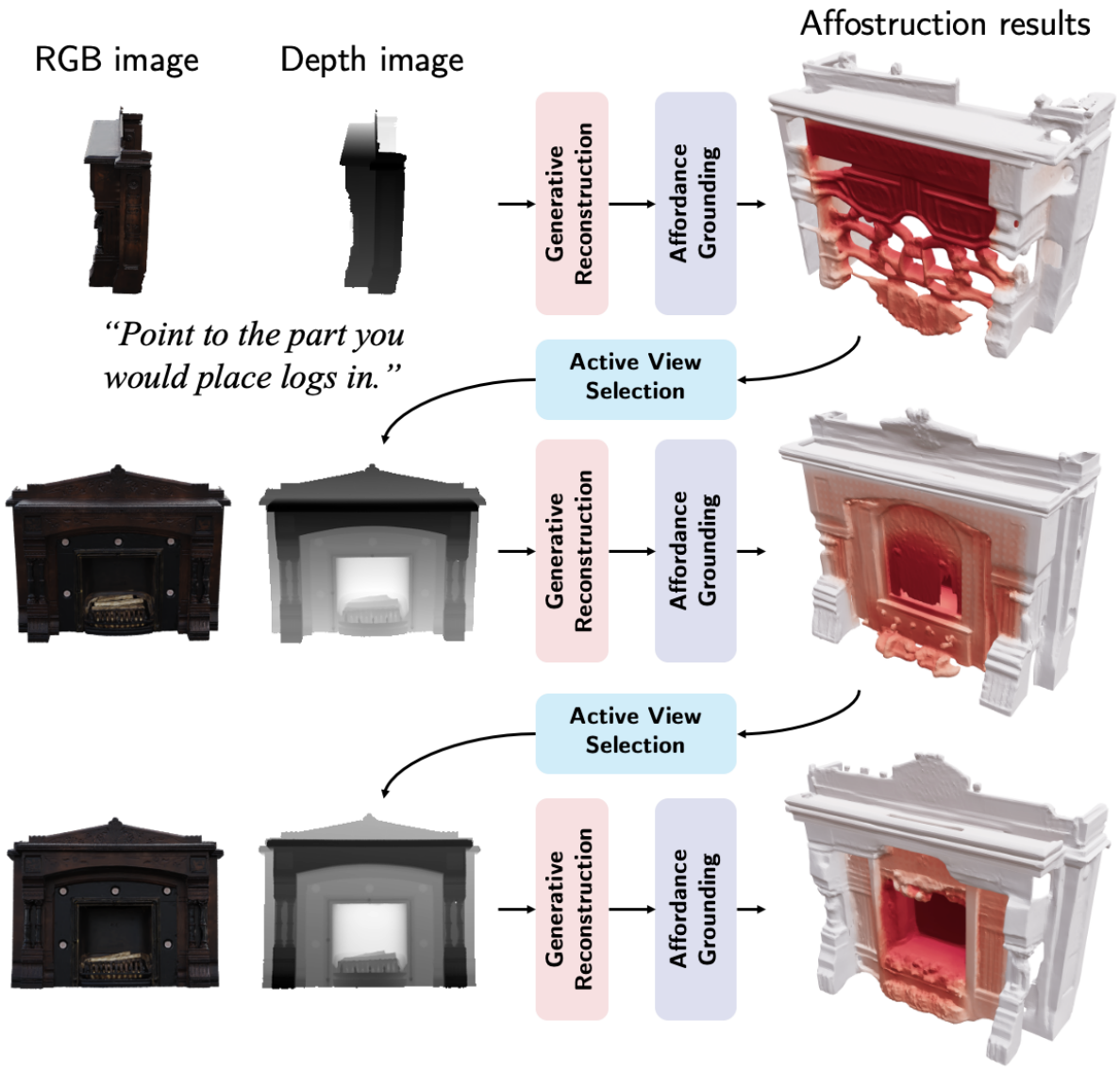

Figure A2. **Progressive improvement through active view selection.** Starting from challenging viewpoints where target areas are barely visible, Affostruction progressively improves through iterative steps: (1) generative reconstruction extrapolates complete structure from partial observations, (2) affordance prediction localizes functional regions on the reconstructed geometry, and (3) active view selection targets informative viewpoints based on predicted affordances. As more views are accumulated through multi-view fusion, both reconstruction quality and affordance localization improve. Only the selected view is shown for clarity.

References

[1] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[2] Junha Lee, Eunha Park, Chunghyun Park, Dahyun Kang, and Minsu Cho. Affogato: Learning open-vocabulary affordance grounding with automated data generation at scale. *arXiv preprint arXiv:2506.12009*, 2025. 1

[3] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1

[7] Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[8] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 3