Mask-Guided High-Fidelity 3D Product Reconstruction via SAM2 and 3D-Consistent Gaussian Splatting Refinement

Zidong Song[1*]
songzd2024@shanghaitech.edu.cn

Boyang Zhou[1*]
zhouby2024@shanghaitech.edu.cn

Zian Chen[1*]
chenza2024@shanghaitech.edu.cn

[1]ShanghaiTech University

Abstract

Creating clean 3D object assets from casual monocular videos remains challenging due to cluttered backgrounds and reconstruction artifacts. We present MOF3R, a segmentation-guided 3D reconstruction pipeline built upon 3D Gaussian Splatting (3DGS). Object masks generated by SAM2 are used as foreground priors through a Mask-Guided Compound Loss that focuses optimization on target objects while suppressing background fitting. We further introduce a geometry-aware pruning framework that removes floating Gaussians and boundary artifacts using multi-view mask consistency, density analysis, and anisotropy constraints. Experiments on the CO3Dv2 dataset show that MOF3R produces cleaner reconstructions with fewer artifacts and sharper object boundaries than original 3DGS, demonstrating the effectiveness of combining semantic guidance with geometric refinement for object-centric 3D reconstruction.

1 Introduction

The acquisition of high-quality 3D digital assets is important for e-commerce, virtual reality, and the creation of digital content. Recent advances in 3D Gaussian Splatting (3DGS) have significantly improved reconstruction efficiency over Neural Radiance Fields (NeRF), enabling fast training and real-time rendering while maintaining high visual quality.

Despite these advantages, applying 3DGS to casually captured mobile videos remains challenging. Such videos often contain cluttered backgrounds, limited viewpoints, and image degradations, causing Gaussian primitives to be allocated to irrelevant regions. Consequently, reconstructions frequently suffer from floating artifacts, noisy structures, and inaccurate object boundaries, reducing asset quality and usability.

Recent progress in video object segmentation provides a promising solution through semantic foreground priors. In particular, SAM2 can generate reliable object masks in video sequences. However, existing reconstruction pipelines typically use segmentation only as a preprocessing step and do not fully exploit semantic information during optimization and geometric refinement.

To address this limitation, we propose Masked Object Fidelity 3D Reconstruction (MOF3R), an object-centric reconstruction framework for monocular videos. The foreground masks generated by SAM2 are incorporated into 3DGS optimization through a Mask-Guided Compound Loss that focuses on the surveillance of target objects and suppresses background fitting. We further introduce a geometry-aware pruning strategy based on multi-view consistency, local density analysis, and anisotropy constraints to remove floating artifacts and refine object boundaries.

*Co-first authors.

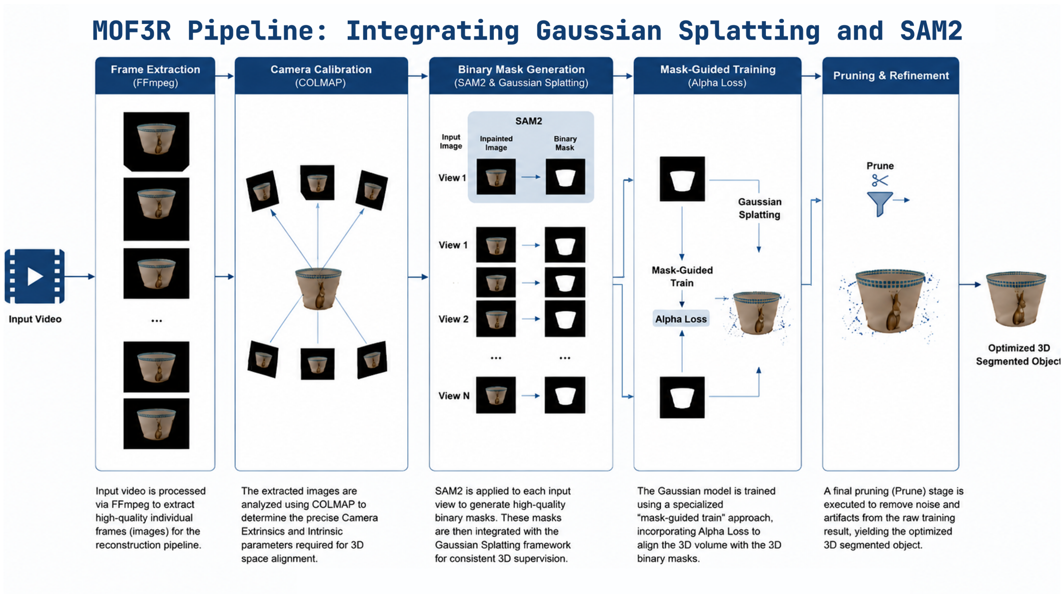

Figure 1: MOF3R: Overall pipeline schematic

The main contributions of this work are summarized as follows:

1. We present a segmentation-guided object-centric 3DGS reconstruction framework for generating high-quality 3D assets from casually captured videos.
2. We introduce a Mask-Guided Compound Loss that incorporates foreground semantic priors into 3DGS optimization, reducing background interference, and improving reconstruction quality.
3. We propose a geometry-aware pruning strategy that combines multi-view consistency, local geometric structure, and anisotropy analysis to remove floating artifacts while preserving fine details.

Experiments on the CO3Dv2 data set demonstrate that the proposed method produces cleaner reconstructions and sharper object boundaries than the original 3DGS.

2 Method

2.1 Overview

Our goal is to reconstruct a high-fidelity object-centric 3D representation from a monocular video. Given an input video sequence $V = I_{i_{i=1}}^{N}$, we first estimate camera parameters $C = K_i, T_{i_{i=1}}^{N}$ using COLMAP and obtain foreground masks $M = M_{i_{i=1}}^{N}$ with SAM2.

Building upon these semantic priors, we propose a reconstruction framework that combines mask-guided optimization with geometry-aware refinement. During training, foreground masks are incorporated into the 3D Gaussian Splatting (3DGS) optimization process to suppress background fitting. After reconstruction, we apply a multi-metric pruning strategy based on multi-view consistency and local geometric properties to remove floating artifacts and refine object boundaries. The overall pipeline is illustrated in Figure 1.

2.2 Mask-Guided Optimization

A major challenge of object-centric reconstruction from unconstrained videos is that standard 3DGS optimizes all image pixels equally, causing Gaussian primitives to fit both the target object and surrounding background regions.

To focus optimization on the object of interest, we incorporate foreground masks generated by SAM2 into the training objective. Given a ground-truth image I, rendered image $\hat{I}$, and binary mask M, we define a mask-constrained L1 loss as

$$\mathcal{L}_{L1}^{mask} = |M \odot (I - \hat{I})|_1$$

where $\odot$ denotes multiplication by element. This formulation restricts photometric supervision to foreground regions and suppresses background fitting during the early stages of optimization.

Directly applying structural similarity loss only inside the mask may introduce instability near object boundaries. Therefore, we construct a composite image

$$I_{comp} = M \odot \hat{I} + (1 - M) \odot I$$

where rendered foreground pixels are blended with the original background. The SSIM loss is then computed on the composite image:

$$\mathcal{L}_{\text{SSIM}}^{mask} = 1 - \text{SSIM}(I, I_{comp})$$

The final training objective is defined as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{L1}^{mask} + \lambda_2 \mathcal{L}_{\text{SSIM}}^{mask}$$

This compound objective encourages Gaussian primitives to concentrate on the target object while preserving global image consistency, resulting in cleaner reconstructions and fewer background artifacts.

2.3 Geometry-Aware Multi-Metric Pruning

Although mask-guided optimization effectively suppresses most background interference, the reconstructed model may still contain floating Gaussians and elongated boundary artifacts caused by limited viewpoints and optimization errors. To further improve reconstruction quality, we introduce a geometry-aware multi-metric pruning strategy.

2.3.1 Multi-view Mask Consistency

For each Gaussian center g_j, we project its 3D position into all visible camera views and examine whether the projected point lies within the corresponding foreground masks.

Gaussians that consistently fall outside foreground regions across multiple valid views are considered background artifacts and removed. To avoid incorrect pruning caused by occlusions, only visible views participate in the voting process.

This consistency check effectively eliminates isolated floating Gaussians around the reconstructed object.

2.3.2 Density and Anisotropy Filtering

Multi-view consistency alone is insufficient to remove all boundary artifacts. Therefore, we further analyze the local geometric structure of each Gaussian using a K-nearest-neighbor (KNN) neighborhood.

For each Gaussian primitive, we estimate its local density and anisotropy level from neighboring Gaussians. A Gaussian is marked as an outlier if it simultaneously exhibits:

1. significantly lower local density than its neighborhood;
2. excessive anisotropy beyond a predefined threshold.

Such primitives typically correspond to stretched structures and trailing artifacts near object boundaries. Removing them leads to a cleaner and more compact reconstruction.

2.3.3 Adaptive Shrinkage

Directly removing all boundary candidates may damage fine structures. To preserve delicate geometric details, we introduce an adaptive shrinkage mechanism.

Instead of immediately deleting uncertain boundary Gaussians, we gradually reduce their scale and opacity:

$$s' = \alpha s, \qquad o' = \beta o$$

where $0 < \alpha, \beta < 1$.

This progressive refinement smooths object boundaries while retaining high-confidence geometric structures, resulting in a cleaner and visually sharper reconstruction.

3 Experiments

3.1 Experimental Setup

We evaluate our method on the CO3Dv2 dataset, a large-scale benchmark containing multi-view video captures of real-world objects under diverse viewpoints and backgrounds. To stress-test reconstruction quality in challenging scenarios, we focus on categories such as Bowl and Teddy, which contain thin-walled structures, self-occlusions, and complex geometries.

3.2 Preprocessing

For each video sequence, frames are extracted at a fixed frame rate and camera intrinsics/extrinsics are estimated using COLMAP. Foreground object masks are generated with SAM2 using a single user-provided prompt on the first frame.

3.3 Training Details

We build our method on top of the original 3D Gaussian Splatting (3DGS) framework. During optimization, we apply the proposed Mask-Guided Compound Loss, which combines mask-constrained L1 supervision with a composite-image SSIM loss. The loss weights λ_1 and λ_2 are empirically set to balance photometric accuracy and structural consistency.

After reconstruction, we apply the proposed Geometry-Aware Multi-Metric Pruning procedure, including multi-view mask consistency checking, KNN-based density and anisotropy filtering, and adaptive shrinkage refinement.

3.4 Baseline and Metrics

We compare our method against original 3D Gaussian Splatting (3DGS) trained directly on the original video frames without segmentation guidance.

We report the following metrics to evaluate reconstruction quality:

- PSNR (Peak Signal-to-Noise Ratio)
- SSIM (Structural Similarity Index)
- LPIPS (Learned Perceptual Image Patch Similarity)

In addition, we provide qualitative comparisons focusing on background suppression, boundary sharpness, and the removal of floating artifacts.

3.5 Quantitative Results

Table 1 reports quantitative results on representative CO3Dv2 sequences.

Figure 2 visualizes the quantitative performance in different object sequences. The proposed pipeline consistently improves perceptual quality while maintaining high structural similarity.

Table 1: Quantitative comparison on CO3Dv2 sequences. Metrics are evaluated within foreground mask regions.

Method	PSNR↑	SSIM↑	LPIPS↓
Original 3DGS	19.07	0.592	0.629
Mask-Guided	21.03	0.931	0.125
MOF3R (ours)	**23.56**	**0.935**	**0.120**

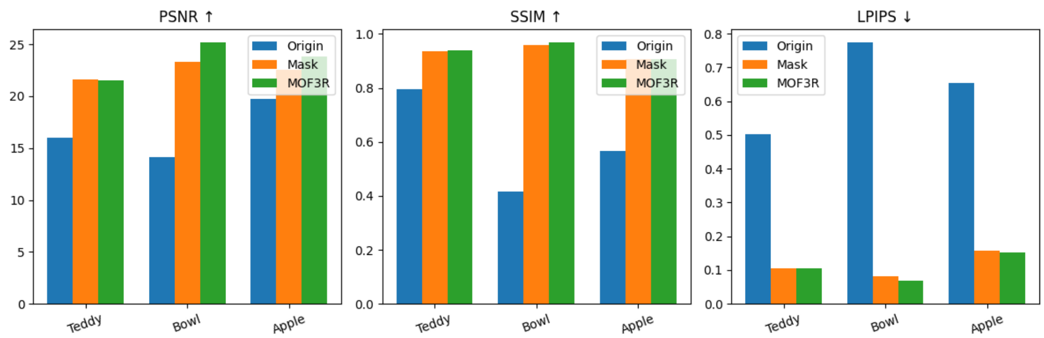

Figure 2: Comparison of reconstruction quality across representative CO3Dv2 sequences.

Compared with original 3DGS, our full pipeline improves the average PSNR from 19.07 dB to 23.56 dB and reduces LPIPS from 0.629 to 0.120. These improvements indicate that semantic guidance effectively suppresses background fitting while geometric pruning removes residual artifacts.

The proposed geometry-aware pruning stage further improves reconstruction quality by removing residual floating artifacts and elongated Gaussian primitives. The improvements are particularly evident on challenging objects such as Bowl, where the full pipeline improves PSNR from 14.10 dB to 25.17 dB while reducing LPIPS from 0.774 to 0.068.

Full per-sequence quantitative results, including different pruning configurations and training iterations, are provided in Appendix A.

3.6 Qualitative Results

Figure 3 presents qualitative comparisons on representative CO3Dv2 scenes. Qualitative comparisons further demonstrate cleaner object boundaries, fewer background artifacts, and more compact Gaussian distributions, validating the complementary benefits of semantic guidance and geometric refinement.

Without foreground supervision, original 3DGS tends to allocate Gaussian primitives to background regions, resulting in severe floating artifacts and blurred object boundaries.

By introducing the proposed mask-guided loss, the optimization process becomes object-centric and substantially suppresses background interference. However, a small number of residual artifacts remain around object boundaries.

The proposed geometry-aware pruning stage further removes isolated and elongated Gaussian primitives, producing cleaner silhouettes and more compact reconstructions. The improvement is particularly visible on the TeddyBear, Bowl, and Apple scenes, where boundary sharpness and visual consistency are significantly enhanced.

3.7 Ablation Study

We conduct an ablation study to evaluate the contribution of each component in the proposed pipeline.

Starting from original 3DGS, we first introduce the mask-guided compound loss, which restricts optimization to foreground regions. We then apply the proposed geometry-aware pruning strategy to remove residual artifacts and elongated Gaussian primitives.

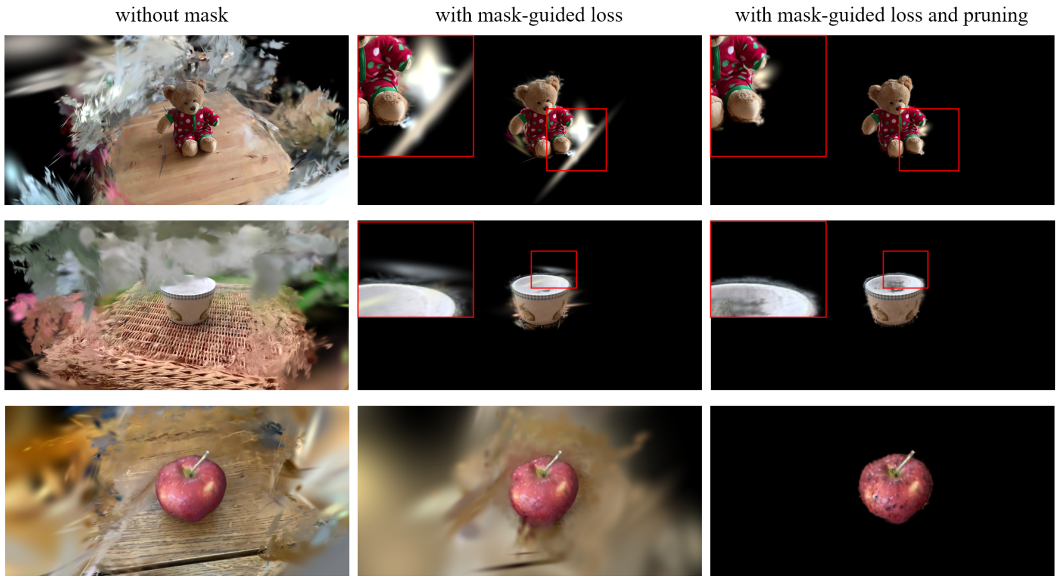

Figure 3: Qualitative comparison of reconstruction results. From left to right: original 3DGS, reconstruction with the proposed mask-guided loss, and the full MOF3R pipeline with geometry-aware pruning. The mask-guided loss effectively suppresses background interference, while the pruning stage further removes floating artifacts and improves object boundary quality.

Table 2: Effect of different pruning strategies.

Pruning Type	PSNR↑	SSIM↑	LPIPS↓
None	21.03	0.931	0.125
Conservative	23.18	0.933	0.123
Reasonable	**23.56**	**0.935**	**0.120**
Aggressive	23.42	0.932	0.121

Results in Table 2 show that the pruning stage further improves reconstruction quality by removing floating artifacts and refining object boundaries where Reasonable pruning showcases the best performance.

4 Conclusion

In this work, we presented MOF3R, a segmentation-guided framework for object-centric 3D reconstruction from monocular videos. By incorporating SAM2-generated foreground masks into the 3D Gaussian Splatting optimization process, our Mask-Guided Compound Loss encourages Gaussian primitives to focus on the target object while suppressing background fitting. We further introduced a geometry-aware multi-metric pruning strategy that combines multi-view mask consistency, local density analysis, and anisotropy filtering to remove floating artifacts and refine object boundaries.

Experiments on challenging categories from the CO3Dv2 dataset demonstrate that the proposed approach produces cleaner reconstructions with fewer background artifacts and sharper object boundaries compared with original 3DGS. The results suggest that semantic priors and geometric refinement provide complementary benefits for object-centric reconstruction in unconstrained environments.

For future work, we plan to investigate tighter integration between segmentation and reconstruction, as well as more robust geometric regularization techniques for handling transparent objects, reflective surfaces, and highly occluded scenes.

References

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL `https://arxiv.org/abs/2408.00714`.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024.

Subhankar Mishra. Clean-gs: Semantic mask-guided pruning for 3d gaussian splatting, 2026. URL `https://arxiv.org/abs/2601.00913`.

Hanqing Wang, Jiaolong Yang, Wei Liang, and Xin Tong. Deep single-view 3d object reconstruction with visual hull embedding, 2018. URL `https://arxiv.org/abs/1809.03451`.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes, 2024. URL `https://arxiv.org/abs/2312.00732`.

A Full Quantitative Results

This appendix provides the complete quantitative results for all evaluated sequences, training iterations, and pruning configurations.

group	scene id	iteration lable	psnr-mean	ssim-mean	lpips-mean
origin	teddy	1000	15.957719	0.795733	0.502974
origin	teddy	3000	18.866552	0.833170	0.392285
origin	bowl	1000	14.101571	0.415512	0.773754
origin	bowl	3000	16.563001	0.589804	0.452322
origin	apple	1000	19.714700	0.566154	0.653520
origin	apple	3000	26.401725	0.809528	0.210779
mask	teddy	1000	15.014329	0.844903	0.262621
mask	teddy	3000	21.628734	0.937388	0.105317
mask	bowl	1000	19.890134	0.929169	0.127335
mask	bowl	3000	23.286411	0.959550	0.082352
mask	apple	1000	16.627887	0.781926	0.312548
mask	apple	3000	22.528242	0.905894	0.157318
MOF3R	teddy	1000	12.385004	0.743488	0.426430
MOF3R	teddy	1000-mask-pruned	12.493246	0.746858	0.422140
MOF3R	teddy	1000-mask-pruned-aggressive	12.659274	0.746094	0.416255
MOF3R	teddy	1000-mask-pruned-conservative	12.586952	0.745608	0.422369
MOF3R	teddy	1000-mask-pruned-extreme	12.659274	0.746094	0.416255
MOF3R	teddy	1000-mask-pruned-reasonable	12.654369	0.747885	0.416781
MOF3R	teddy	3000	19.715921	0.921837	0.153458
MOF3R	teddy	3000-mask-pruned	20.170774	0.924714	0.149210
MOF3R	teddy	3000-mask-pruned-aggressive	21.969686	0.927167	0.137814
MOF3R	teddy	3000-mask-pruned-conservative	21.317001	0.929250	0.140480
MOF3R	teddy	3000-mask-pruned-extreme	21.969686	0.927167	0.137814
MOF3R	teddy	3000-mask-pruned-reasonable	21.715209	0.930042	0.137910
MOF3R	bowl	1000	20.116411	0.931011	0.126487
MOF3R	bowl	1000-mask-pruned	20.264461	0.933845	0.122918
MOF3R	bowl	1000-mask-pruned-aggressive	22.358688	0.953486	0.091790
MOF3R	bowl	1000-mask-pruned-conservative	21.520443	0.946677	0.103117
MOF3R	bowl	1000-mask-pruned-extreme	22.358688	0.953486	0.091790
MOF3R	bowl	1000-mask-pruned-reasonable	21.783171	0.948651	0.099945
MOF3R	bowl	3000	23.339291	0.959306	0.082510
MOF3R	bowl	3000-mask-pruned	23.688310	0.961111	0.079778
MOF3R	bowl	3000-mask-pruned-aggressive	25.359025	0.967092	0.068848
MOF3R	bowl	3000-mask-pruned-conservative	24.791282	0.966737	0.070560
MOF3R	bowl	3000-mask-pruned-extreme	25.359025	0.967092	0.068848
MOF3R	bowl	3000-mask-pruned-reasonable	25.170994	0.967742	0.068185
MOF3R	apple	1000	16.801979	0.786755	0.300191
MOF3R	apple	1000-mask-pruned	17.028800	0.794986	0.294224
MOF3R	apple	1000-mask-pruned-aggressive	17.036958	0.774801	0.316318
MOF3R	apple	1000-mask-pruned-conservative	17.145694	0.791379	0.299683
MOF3R	apple	1000-mask-pruned-extreme	17.036958	0.774801	0.316318
MOF3R	apple	1000-mask-pruned-reasonable	17.171169	0.785992	0.306671
MOF3R	apple	3000	22.499137	0.905678	0.157529
MOF3R	apple	3000-mask-pruned	22.727309	0.907408	0.155726
MOF3R	apple	3000-mask-pruned-aggressive	23.708253	0.900295	0.156291
MOF3R	apple	3000-mask-pruned-conservative	23.429492	0.906389	0.156643
MOF3R	apple	3000-mask-pruned-extreme	23.708253	0.900295	0.156291
MOF3R	apple	3000-mask-pruned-reasonable	23.783064	0.906305	0.153110

Table 3: The Result of all Metrics.

Iteration Label Definition. The *iteration label* column indicates the reconstruction stage at which the evaluation was performed:

- **1000**, **3000**: reconstruction results obtained after 1,000 and 3,000 optimization iterations, respectively.
- **1000-mask-pruned**, **3000-mask-pruned**: results after applying the proposed geometry-aware pruning procedure to the corresponding reconstruction checkpoint.
- **1000-mask-pruned-conservative**: pruning using relatively loose thresholds to preserve more Gaussian primitives and fine structures.
- **1000-mask-pruned-reasonable**: the default pruning configuration used in our final pipeline, balancing artifact removal and structure preservation.
- **1000-mask-pruned-aggressive**: pruning using stricter thresholds to remove a larger number of potentially noisy Gaussian primitives.
- **1000-mask-pruned-extreme**: an even more aggressive pruning configuration designed to investigate the upper limit of artifact removal.

The same naming convention applies to the corresponding 3,000-iteration checkpoints.