

GaussianZoom: Progressive Zoom-in Generative 3D Gaussian Splatting with Geometric and Semantic Guidance

Supplementary Material

This supplementary material is organized as follows. Section A presents the pseudocode of the GaussianZoom pipeline. Section B elaborates on the formulation of our continuous Level-of-Detail representation. Additional qualitative comparisons are provided in Section C. Section D reports additional quantitative results, including ablation studies of the LoD and VLM components, comparisons under arbitrary magnification factors, and a runtime and memory efficiency analysis. Limitation discussions are provided in Section E.

A. Pseudocode

Algorithm A shows the pseudocode for GaussianZoom pipeline.

Algorithm A GaussianZoom Pipeline

Require: Input images $\{I_i\}_{i=1}^N$, camera poses $\{P_i\}_{i=1}^N$

- 1: **Train** coarse 3DGS \mathcal{G}_0 on $(\{I_i\}, \{P_i\})$
- 2: **for** $t = 1$ to N **do** ▷ zoom step
- 3: Generate zoom-in cameras $\{P_{t,i}^z\}$ by scaling focal length
- 4: Render wide-view image R_t^w and zoom-in RGBD maps $\{(R_{t,i}^z, D_{t,i}^z)\}$ from \mathcal{G}_{t-1}
- 5: **for** each zoom-in view i **do**
- 6: $c_{t,i} \leftarrow \text{VLM}(R_t^w, R_{t,i}^z)$
- 7: $j \leftarrow \text{NearestView}(P_{t,i}^z)$
- 8: $\mathbf{F}_{t,i}^z \leftarrow \text{Enc}(R_{t,i}^z)$
- 9: $\tilde{\mathbf{F}}_{t,i} \leftarrow W_{j \rightarrow i}(\mathbf{F}_{t,j}, D_{t,i}^z)$
- 10: $I_{t,i}^{\text{SR}} \leftarrow \mathcal{S}(\tilde{\mathbf{F}}_{t,i}, \mathbf{F}_{t,i}^z, c_{t,i})$
- 11: **end for**
- 12: **Train** 3DGS at scale t on $\{I_{t,i}^{\text{SR}}\}$ to obtain \mathcal{G}_t
- 13: **end for**

B. Continuous Level-of-Detail Formulation

B.1. Densification

During each zoom step, primitives from previous steps are treated as fixed: their parameters remain unchanged, and only their screen-space gradients are used for densification.

For splitting operations, two cases are considered. For a fixed primitive, two scaled-down child primitives are directly appended beneath it as its children. For a non-fixed primitive, the original node is removed and replaced by two scaled children under the original primitive’s parent, avoiding redundant representations at the same level.

For cloning operations, a newly instantiated primitive with identical attributes is inserted as an additional child under the parent primitive.

B.2. Opacity Adjustment

In zoom-in setting where the focal length changes from f to f' , the scale projection coefficient of a primitive is defined as $\psi = d/f$ and becomes $\psi' = d'/f'$ after zoom-in. Since the camera pose is nearly unchanged, the distance variation is negligible, i.e., $d' \approx d$, giving

$$\frac{\psi'}{\psi} \approx \frac{f}{f'} \quad (1)$$

showing that the change of scale is dominated by the focal length modification.

The ratio determines whether a primitive should rely on its parent or child in the LoD hierarchy. When $\psi'/\psi > 1$, the new viewpoint corresponds to a coarser effective scale, and the parent primitive becomes the appropriate representation; when $\psi'/\psi < 1$, the viewpoint demands finer detail and the child primitive becomes preferable. If neither condition holds or if the primitive lacks a valid parent/child counterpart, no transition is needed and opacity adjustment is skipped entirely.

When a transition is required, opacity must change smoothly to avoid popping artifacts. Let ψ_p and ψ_c denote the stored LoD ratios of the parent and child primitives, respectively. Because the parent and child lie close in space, their distances satisfy $d_p \approx d_c \approx d'$, yielding

$$\frac{\psi'}{\psi_p} \approx \frac{f_p}{f'}, \quad \frac{\psi'}{\psi_c} \approx \frac{f_c}{f'}, \quad \frac{f_p}{f_c} \approx \frac{1}{s} \quad (2)$$

We adjust the opacity using a logarithmic interpolation with base of zoom factor s :

$$w_p = 1 - \log_s\left(\frac{\psi'}{\psi_p}\right), \quad w_c = 1 + \log_s\left(\frac{\psi'}{\psi_c}\right). \quad (3)$$

The parent weight decreases as the zoom-in requires more detail, while the child weight increases correspondingly, and vice versa for zoom-out, forming a symmetric adjustment around the transition point. The normalization of these weights follows from the reciprocal structure of the ratios. Because $d_p \approx d_c$, we have

$$\log_s\left(\frac{\psi'}{\psi_c}\right) - \log_s\left(\frac{\psi'}{\psi_p}\right) \approx \log_s\left(\frac{f_p}{f_c}\right) \approx -1 \quad (4)$$

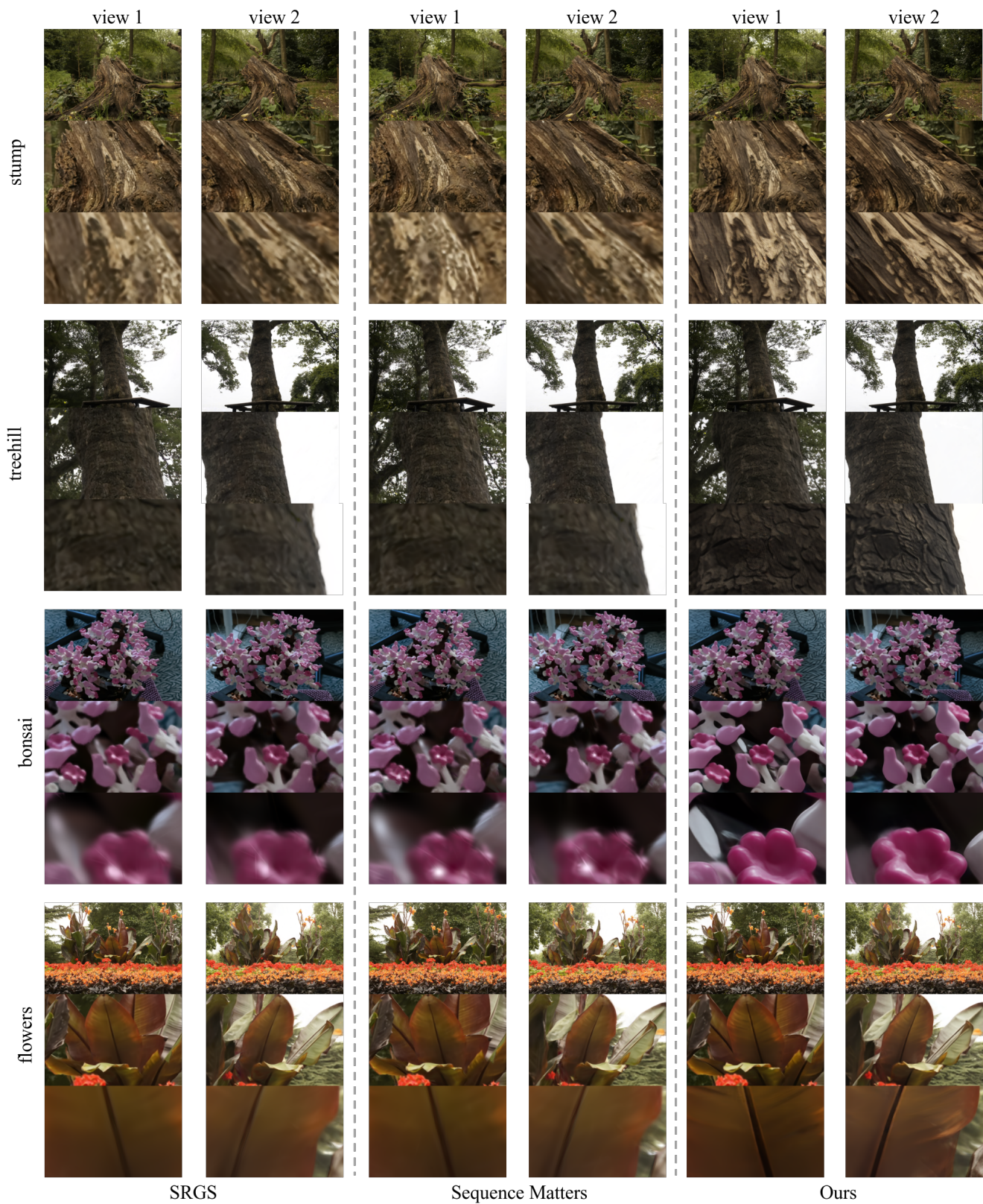


Figure A. Additional qualitative results of zoom-in task on Mip-NeRF360 dataset.



Figure B. Situations with larger magnification. Text below images are prompts generated by vision-language model.

Method	CLIP-IQA \uparrow			MUSIQ \uparrow			NIQE \downarrow		
	11.3 \times	20.4 \times	46.7 \times	11.3 \times	20.4 \times	46.7 \times	11.3 \times	20.4 \times	46.7 \times
SRGS [2]	0.322	0.252	0.258	46.77	25.72	17.65	6.648	9.768	13.28
Sequence Matters [3]	0.319	0.251	0.251	47.15	26.23	17.55	6.083	9.292	13.26
Ours	0.370	0.334	0.397	55.39	43.82	36.28	5.092	5.479	6.187

Table A. Quantitative comparison under arbitrary magnification on Mip-NeRF360. Our progressive pipeline maintains high perceptual quality across all tested magnification levels, while competing methods degrade significantly at higher zoom factors.

which yields the identity

$$w_p + w_c \approx 1, \quad (5)$$

Thus, opacity is conserved during the transition, ensuring that the parent-child mixture produces a visually stable and alias-free LoD change across zoom levels.

A detailed opacity adjustment is shown in Algorithm B.

Algorithm B Opacity Adjustment

Require: Ratio of scale projection coefficient $r = \frac{\psi'}{\psi}$

- 1: **if** $r \in [1, s)$ and has parent **then**
 - 2: $w = 1 - \log_s r$
 - 3: **else if** $r \in (\frac{1}{s}, 1)$ and has child **then**
 - 4: $w = 1 - \log_s r$
 - 5: **else if** $r > 1$ and has no parent **then**
 - 6: $w = 1$
 - 7: **else if** $r < \frac{1}{s}$ and has no child **then**
 - 8: $w = 1$
 - 9: **else**
 - 10: $w = 0$
 - 11: **end if**
-

C. Additional Qualitative Results

We further show additional qualitative results of zoom-in task on the Mip-NeRF360 [1] dataset in Fig. A. As the zoom factor increases, competing methods tend to produce blurrier outputs with reduced semantic detail. In contrast, our approach maintains clearer structures and semantically richer textures across zoom levels, demonstrating improved quality under zoom-in tasks.

D. Additional Quantitative Results

D.1. Ablation of LoD and VLM Components

As noted in the main paper, the benefits of the LoD hierarchy and VLM-guided prompting become more pronounced at higher magnifications. Table B presents quantitative ablation results under 64 \times magnification using no-reference image quality metrics. Both components contribute positively, with the full model achieving the best scores across CLIP-IQA, MUSIQ, and NIQE.

Method	CLIP-IQA \uparrow	MUSIQ \uparrow	NIQE \downarrow
Ours w/o LoD	0.319	37.44	9.18
Ours w/o VLM	0.431	38.11	5.84
Ours	0.436	42.21	5.53

Table B. Quantitative ablation of LoD and VLM components at 64 \times magnification. Higher CLIP-IQA/MUSIQ and lower NIQE indicate better perceptual quality.

D.2. Results under Arbitrary Magnification

A key advantage of our LoD over prior p approaches is its ability to support arbitrary magnification factors beyond 4 \times . Table A reports no-reference quality metrics for three representative magnification levels (11.3 \times , 20.4 \times , 46.7 \times) on the Mip-NeRF360 dataset, comparing against SRGS [2] and Sequence Matters [3]. Our method consistently outperforms both baselines at all magnification levels, demonstrating robustness to increasing zoom factors.

D.3. Runtime and Memory Efficiency

Table C reports the average per-image super-resolution time and peak GPU memory when upsampling to 1K resolution under identical hardware.

Method	RunTime (s) ↓	Memory (GB) ↓
SRGS [2]	1.54	3.56
Sequence Matters [3]	6.08	23.48
Ours	2.57	21.72

Table C. Per-image SR runtime and peak GPU memory when up-sampling to 1K resolution under identical hardware conditions.

E. Limitation

Our method will encounter difficulties at very high magnifications, where current vision-language models struggle to infer coherent structures, leading to semantically weak textures. As shown in Fig. B, when the zoom range becomes extreme, current vision-language models often fail to correctly identify the target region, producing text prompts that no longer correspond to the underlying scene content.

References

- [1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 3
- [2] Xiang Feng, Yongbo He, Yubo Wang, Yan Yang, Wen Li, Yifei Chen, Zhenzhong Kuang, Jianping Fan, Yu Jun, et al. Srgs: Super-resolution 3d gaussian splatting. *arXiv preprint arXiv:2404.10318*, 2024. 3, 4
- [3] Hyun-kyu Ko, Dongheok Park, Youngin Park, Byeonghyeon Lee, Juhee Han, and Eunbyung Park. Sequence matters: Harnessing video models in 3d super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4356–4364, 2025. 3, 4