PBE: A PHYSICS-GUIDED BETA-PRIMITIVE EXTENSION FOR FEED-FORWARD 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

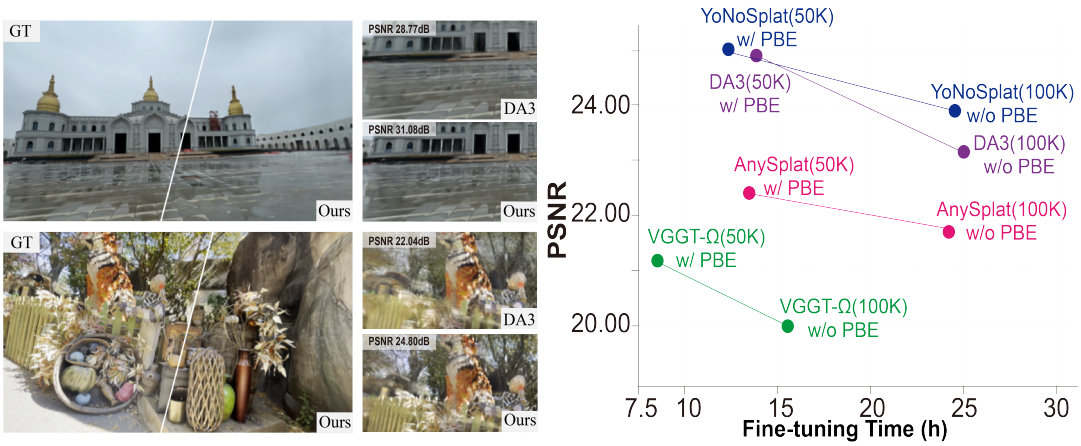


Figure 1: **Rendering quality and training efficiency of PBE. Left:** Qualitative comparisons on mirror reflections (top) and specular highlights (bottom). PBE better preserves reflected structures and fine details.**Right:** In the 16-view setting, PBE achieves higher PSNR across four models with less training time.

ABSTRACT

Feed-forward Gaussian reconstruction predicts renderable 3D scenes in a single pass, but its Gaussian kernels and low-order spherical harmonics offer limited flexibility in spatial shape and view-dependent appearance. Existing Beta kernels and Spherical Beta (SB) functions offer greater flexibility through adjustable spatial decay and localized angular responses. However, predicting Beta/SB parameters in a feed-forward manner is challenging because geometry and appearance require different cues, and sparse observations do not fully constrain view-dependent appearance. We propose **PBE**, a **P**hysics-guided **B**eta-primitive **E**xtension that addresses this problem through input-dependent base construction and residual prediction. PBE constructs single-view Beta/SB bases from predicted depth and camera parameters, and multi-view bases from depth-consistent support-view observations. Attribute-specific gates combine these bases, while learned residuals correct inaccurate estimates. PBE integrates into the primitive-parameter prediction stage of existing reconstruction models while preserving their feature backbones and single-pass inference. Trained solely on DL3DV, it consistently improves rendering quality on Depth Anything 3, VGGT-Ω, YoNoSplat, and AnySplat under matched epoch budgets. Notably, Depth Anything 3 on Mip-NeRF 360 improves from 24.25 to 26.01 dB after five epochs without additional fine-tuning. Ablations further show that physics-informed Beta bases outperform random bases under the same residual refinement. These results jointly validate the effectiveness of PBE.

1 INTRODUCTION

Reconstructing a renderable 3D scene from a set of images is a fundamental problem in computer vision, with broad applications in augmented reality (Engel et al., 2023), robotics (Zhu et al., 2024),

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.

and autonomous navigation (Mur-Artal et al., 2015). Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) have enabled high-quality rendering through an explicit set of images. Both paradigms generally require scene-specific optimization, which limits their use in applications that demand immediate reconstruction of unseen scenes. Feed-forward Gaussian reconstruction avoids this optimization by predicting scene primitives directly from input images. However, Gaussian kernels have a fixed spatial decay form, while low-order spherical harmonics (SH) have limited capacity for localized view-dependent effects such as sharp specular highlights. Existing Beta kernels and Spherical Beta (SB) functions offer greater flexibility through adjustable spatial decay and localized angular lobes, respectively (Liu et al., 2025), while retaining an explicit splatting pipeline.

Adapting Beta/SB representations to feed-forward prediction requires accounting for the different evidence available for their parameters. Local depth variation and surface geometry provide useful cues for spatial Beta decay and SB lobe directions, whereas appearance also depends on material, illumination, and viewing conditions. Direct regression, even with depth features, leaves these relationships to be learned implicitly, while fixed analytic estimates may over-constrain appearance.

To this end, we propose **PBE**, a **P**hysics-guided **B**eta-primitive **E**xtension for feed-forward Gaussian reconstruction. PBE replaces direct Gaussian/SH attribute regression with explicit Beta/SB base construction and learned residual prediction. Please refer to Table A1 in Appendix A.2 for details. It constructs explicit Beta/SB parameter bases as a *soft physical prior* from predicted geometry and depth-consistent support-view observations, including multi-view color evidence, combines them through attribute-specific gates, and refines them with learned residuals. This design accounts for the different evidence available for spatial and appearance attributes while allowing inaccurate estimates to be corrected. PBE integrates into the primitive-parameter prediction stage, retaining the host model's feature backbone and single-pass inference without per-scene optimization.

Experiments on Depth Anything 3 (DA3), VGGT-Ω, YoNoSplat, and AnySplat show consistent rendering-quality improvements over their corresponding baselines under matched epoch budgets. Notably, after five epochs of training exclusively on DL3DV, PBE improves DA3's PSNR on Mip-NeRF 360 from 24.25 to 26.01 dB without additional fine-tuning. Ablations further show that geometry-informed bases outperform random bases under the same residual refinement scheme, supporting the value of explicit geometric guidance for feed-forward Beta/SB prediction.

Our contributions are summarized as follows:

- **Lightweight integration.** We introduce PBE, a lightweight module with only 0.35M learnable parameters that integrates into existing feed-forward Gaussian reconstruction methods while preserving their single-pass inference.
- **Geometry-informed bases.** We construct renderable spatial Beta bases from geometric information and refine them with learned residuals that can correct inaccurate estimates.
- **Multi-view appearance guidance.** We propose a pose-guided SB base that uses reliable support-view observations to inform localized view-dependent appearance beyond single-view RGB evidence.
- **Cross-architecture validation.** Comparative experiments and ablation studies demonstrate the applicability and effectiveness of PBE across multiple feed-forward reconstruction architectures.

2 Related Work

2.1 3D Gaussian Splatting and Beta Representations

3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) represents a scene with anisotropic Gaussian primitives parameterized by position, opacity, covariance, and SH appearance coefficients. Its explicit representation and differentiable rasterizer enable real-time rendering with high visual quality. Mip-Splatting (Yu et al., 2024) and Multi-Scale 3D Gaussian Splatting (Yan et al., 2024) improve scale consistency and reduce aliasing. N-Dimensional Gaussians (N-DG) (Diolatzis et al., 2024) and 6D Gaussian Splatting (6DGS) (Gao et al., 2025) extend the representation to higher-dimensional spaces and conditional slicing to model more complex view-dependent effects. These studies im-

prove the expressiveness or scale robustness of splatting representations, but they continue to estimate their scene parameters through per-scene optimization.

Deformable Beta Splatting (Liu et al., 2025) replaces the Gaussian kernel with a compact-support, shape-controllable Beta kernel and demonstrates the representational benefits of Beta primitives. PBE builds on this insight but addresses the more under-constrained feed-forward setting. Instead of optimizing a free set of Beta primitives for each scene, our model constructs a geometry-aware Beta base from predicted depth and camera parameters and learns corrective residuals for unseen scenes. It further couples the spatial Beta representation with a multi-lobe SB appearance model and explicit cross-view fitting.

2.2 Feed-Forward 3D Reconstruction

Feed-forward reconstruction methods avoid per-scene optimization by directly inferring scene structure from one or more images. DUSt3R (Wang et al., 2024) reconstructs point maps from image pairs, but extending it to large image collections requires a global alignment stage that adds cost and may introduce instability. Fast3R (Yang et al., 2025) jointly processes large image collections and removes this post-hoc alignment, while Pi3 (Wang et al., 2026b) adopts a permutation-equivariant architecture to process an arbitrary number of views without selecting a fixed reference image. These methods produce strong geometric estimates, but they do not directly provide a complete splatting representation with compact view-dependent appearance.

In contrast, recent studies have extended 3DGS toward feed-forward reconstruction by directly predicting Gaussian representations from input images. Whether developed as dedicated feed-forward 3DGS models, such as DepthSplat (Xu et al. (2025)), AnySplat (Jiang et al. (2025)), and YoNoSplat (Ye et al. (2026)), or constructed by equipping large-scale 3D reconstruction models, such as DA3 (Lin et al. (2025)) and VGGT-Ω (Wang et al. (2026a)), with Gaussian prediction heads, these approaches can generally be abstracted into image feature extraction, cross-view interaction, and Gaussian parameter decoding. Despite their architectural differences, these approaches predominantly treat primitive attributes as quantities to be regressed from learned features. Simple depth-feature fusion does not explicitly encode ray-distance back-projection, surface differential geometry, pose-guided correspondence, or the different observability of geometry and material parameters.

3 Method

3.1 Overview and Representation

Given input images $\mathcal{I}$, PBE uses visual features $\mathbf{F}$, depth maps $\mathcal{D}$, and camera parameters $\mathcal{C}$ from the host model to predict spatial Beta and Spherical Beta (SB) attributes in a single forward pass:

$$\left\{\beta_i, \{\mathbf{a}_{i,\ell}, \boldsymbol{\mu}_{i,\ell}, k_{i,\ell}\}_{\ell=1}^{3}\right\}_{i=1}^{N} = \mathrm{PBE}_\theta(\mathcal{I}, \mathbf{F}, \mathcal{D}, \mathcal{C}), \tag{1}$$

where i indexes the N primitives, ℓ indexes three SB lobes, and θ denotes learnable parameters. Each primitive is associated with an anchor-view pixel. The exponent β_i controls spatial decay, while $\mathbf{a}_{i,\ell}$, $\boldsymbol{\mu}_{i,\ell}$, and $k_{i,\ell}$ specify lobe color, direction, and sharpness. Positions, scales, rotations, opacities, and base colors follow the original prediction branch, together defining a renderable Beta/SB representation (Liu et al., 2025).

As illustrated in Figure 2, Part A of the figure shows the overall prediction pipeline using DA3 as an example, while Part B illustrates the detailed PBE prediction process. The computation procedures are described in Section 3.2 and Section 3.3. PBE constructs single-view bases from geometry and input color, incorporates support-view observations through attribute-specific base fusion, and predicts residual corrections to refine the resulting parameters.

3.2 Geometry-Informed Parameter Bases

PBE converts geometric cues into explicit single-view parameter bases for spatial Beta decay and SB appearance. Geometry guides the spatial decay and lobe directions, while lobe sharpness and color are initialized to provide a starting point for subsequent refinement.

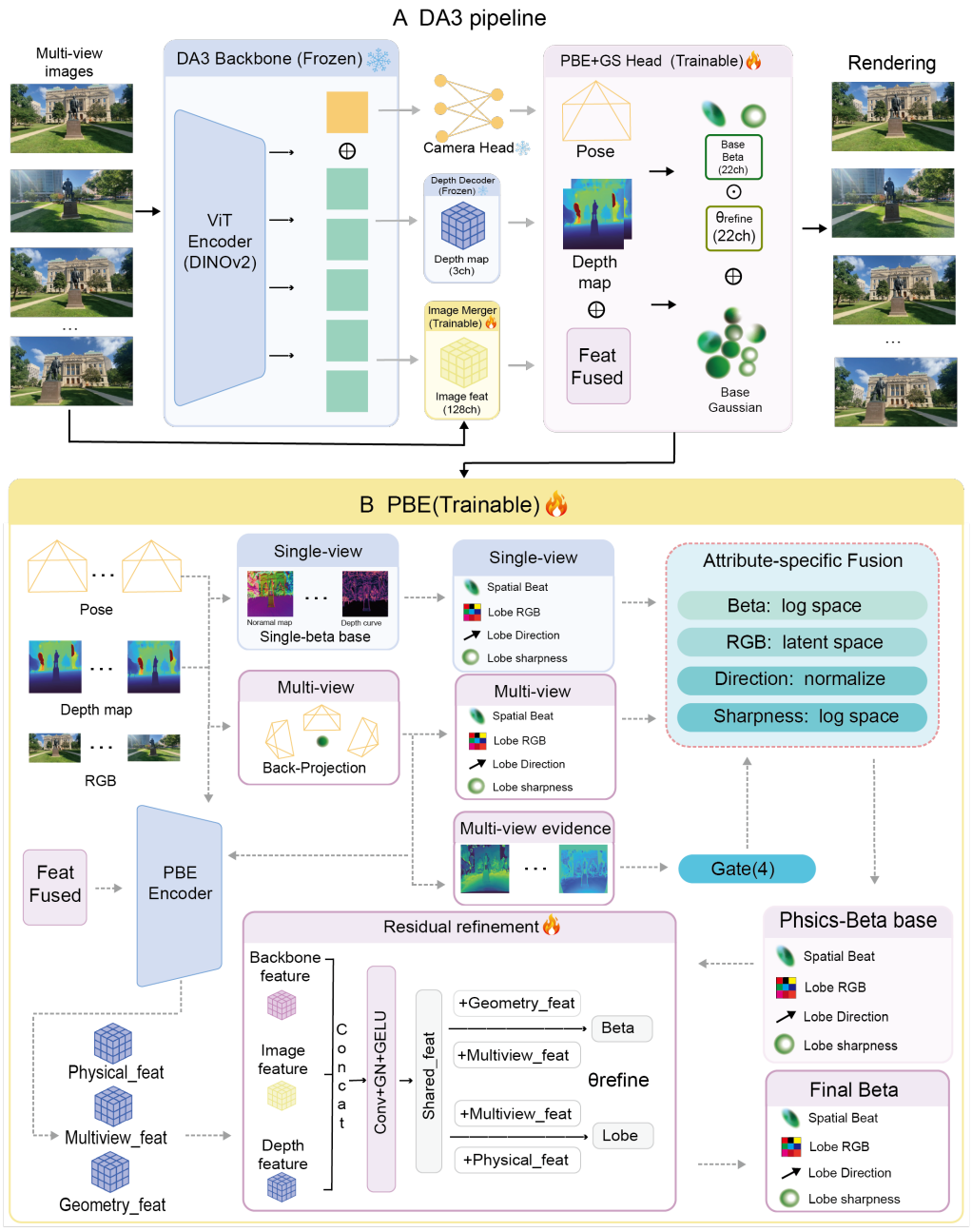


Figure 2: **Architecture of PBE.** PBE derives geometric and multi-view cues from the predicted depth and poses to construct renderable spatial Beta and three-lobe SB bases. Parameter-specific branches then fuse these cues with visual features and predict adaptive base fusion and residual corrections. The resulting Beta/SB attributes extend the original Gaussian primitives.

Geometric cues. We back-project ray-distance depth into camera-space points and estimate camera-facing normals $\mathbf{n}_i$ through local finite differences. Camera rays and normals provide viewing and reflection cues. Median-normalized log depth yields compressed gradients $(g_{x,i}, g_{y,i})$, while local normal variation provides a curvature proxy κ_i. We set $m_i = 1$ when depth is finite and positive and the estimated normal is non-degenerate; otherwise, $m_i = 0$, and fallback bases are used. Numerical thresholds are provided in Appendix A.1. Depth confidence conditions the network without directly scaling the bases; depth, intrinsics, and confidence are detached during cue construction.

Spatial Beta base. We map local geometric variation to a spatial decay reference:

$$G_i = \mathrm{clamp}\left(0.7\frac{\sqrt{g_{x,i}^2 + g_{y,i}^2}}{\sqrt{2}} + 0.3\kappa_i,\ 0,\ 1\right), \tag{2}$$
$$\log \beta_i^{\mathrm{sv}} = (1 - m_i)\log 4 + m_i G_i \log 64.$$

Valid geometry produces exponents in $[1, 64]$, with stronger variation yielding more concentrated responses; invalid geometry falls back to $\beta_i^{\mathrm{sv}} = 4$. This mapping supplies a heuristic reference rather than fixing the final prediction.

Single-view SB base. We construct three directions around $-\mathbf{n}_i$ using a view-conditioned tangent frame, with polar angle $45°$, azimuths $(0°, 120°, 240°)$, and initial sharpness values $(2, 4, 8)$. All SB viewing directions use the camera-to-surface convention. Lobe colors are initialized from the input RGB with scale $\lambda = 0.03$ and reference-response normalization. This scale controls the initial amplitude and is an initialization hyperparameter rather than a physical appearance prior. Invalid geometry uses a camera-ray direction fallback and zero color bases. Construction formulas and numerical handling are provided in Appendix A.1.

3.3 Multi-View Evidence and Attribute-Specific Fusion

Multi-view evidence and bases. For each anchor view, we select up to two other input views with the highest symmetric, depth-consistent overlap. Reprojected points provide the support RGB and ray-distance depth, weighted by projection validity and relative depth agreement. The support fraction s_i, depth and pixel-footprint errors $(e_{d,i}, e_{f,i})$, weighted RGB mean and variance, and reliability $\rho_i = s_i \exp(-e_{d,i}/0.05 - e_{f,i}/0.25)$ form a ten-channel evidence tensor $\mathbf{E}_{\mathrm{mv}}$. The evidence tensor summarizes both geometric reliability and cross-view color statistics. After encoding, its features serve two purposes: predicting attribute-specific gates that combine single-view and multi-view parameter bases, and conditioning residual prediction to refine the fused estimates. The residual feature pathway is not multiplied by these gates, allowing multi-view evidence to inform corrections even when fusion favors the single-view base. The analytical construction of multi-view evidence and parameter bases is performed without gradient tracking; the evidence encoder, the fusion gates, and the residual predictors remain trainable.

For spatial Beta, multi-view reliability modulates the single-view geometric estimate as $\log \beta_i^{\mathrm{mv}} = \log 4 + \rho_i(G_i \log 64 - \log 4)$, moving it toward the reference exponent 4 when the support is weak. For SB, weighted directional moments update geometric lobe proposals using depth consistency, initial lobe responses, and positive color residuals relative to the support-view RGB mean. Each lobe's color is then fitted independently using its updated direction and initial sharpness, followed by a heuristic sharpness update from directional concentration. Colors are not refitted afterward. The support mean does not replace the original base color $\mathbf{c}_0$, and the anchor RGB is excluded from fitting. Detailed estimates are given in Appendix A.1.

Attribute-specific fusion. A 1×1 convolution on the encoded multi-view evidence predicts four gate corrections $\delta_{i,j}$:

$$\alpha_{i,j} = m_i^{\mathrm{mv}}\, \sigma(\mathrm{logit}(\widetilde{\rho}_i) + \delta_{i,j}), \qquad j \in \{\beta, c, d, k\}, \tag{3}$$

where $\widetilde{\rho}_i = \mathrm{clamp}(\rho_i, 10^{-4}, 1 - 10^{-4})$ and m_i^{mv} is set only when the local geometry is valid and the support is positive. Larger weights favor multi-view bases; a zero mask retains single-view bases. Each appearance gate is shared across the three lobes.

For attribute j, fusion takes the form

$$\mathbf{u}_i^{j,\mathrm{base}} = (1 - \alpha_{i,j})\mathbf{u}_i^{j,\mathrm{sv}} + \alpha_{i,j}\mathbf{u}_i^{j,\mathrm{mv}}, \tag{4}$$

where $\mathbf{u}^j$ represents $\log \beta$, inverse-Softplus color latents, direction vectors, or $\log k$, respectively. Appearance fusion is applied separately to each lobe, and blended directions are normalized. These gates control only parameter bases: multi-view features are also added directly to the residual-branch features without gate multiplication.

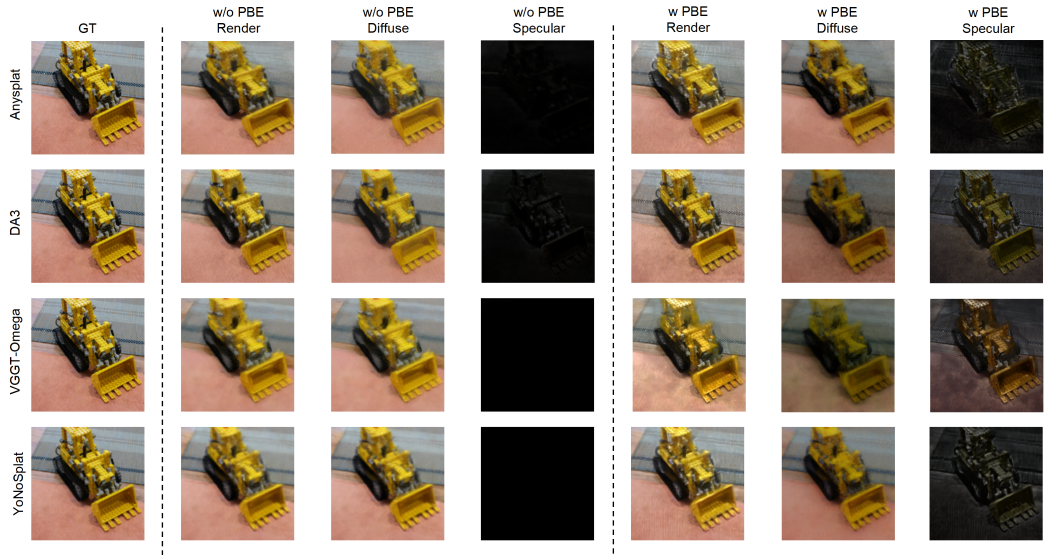


Figure 3: **Light Decomposition.** We compared the performance of PBE and the original baselines on DA3, VGGT-Ω, AnySplat, and YoNoSplat. Our PBE method can effectively decompose diffuse and specular components, as shown in the figure.

3.4 Residual Prediction and Training

Residual prediction. Parameter-specific branches refine the fused bases. Spatial Beta and direction prediction use geometric cues, while color and sharpness primarily use visual and multi-view evidence. The updates are

$$\begin{aligned} \log \beta_i &= \mathrm{clip}\big(\log \beta_i^{\mathrm{base}} + \Delta b_i,\ \log 0.25, \log 256\big)\,, \\ \boldsymbol{\mu}_{i,\ell} &= \mathrm{normalize}_\epsilon\big(\boldsymbol{\mu}_{i,\ell}^{\mathrm{base}} + \Delta \boldsymbol{\mu}_{i,\ell}\big)\,, \\ \log k_{i,\ell} &= \mathrm{clip}\big(\log k_{i,\ell}^{\mathrm{base}} + \Delta h_{i,\ell},\ \log 1, \log 256\big)\,, \\ \mathbf{a}_{i,\ell} &= \phi_\xi\big(\mathcal{E}(\mathbf{a}_{i,\ell}^{\mathrm{base}}) + \Delta \mathbf{a}_{i,\ell}\big)\,. \end{aligned} \tag{5}$$

Here, $\phi_\xi(z) = \xi^{-1}\log(1+e^{\xi z})$ with $\xi = 10\log 2$, and $\mathcal{E}$ is its stabilized inverse encoding. Fused color latents are mapped through ϕ_ξ to obtain color bases, with invalid geometry being assigned zero bases before refinement. Stabilized encoding permits finite residual updates from zero-color bases. Direction normalization uses $\mathbf{x}/\max(\|\mathbf{x}\|_2, 10^{-6})$, and directions are rotated into the world coordinates before rendering. Residuals can depart from the constructed bases within the specified parameter ranges.

Training. We train the extended Gaussian prediction head, including the Beta/SB residual and gate predictors, together with the PBE encoders; all other modules are frozen. The input-view rendering is supervised by

$$\mathcal{L} = \|\widehat{\mathcal{I}} - \mathcal{I}\|_1 + 2\mathcal{L}_{\mathrm{grad}} + \mathcal{L}_{\mathrm{sep}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{grad}}$ matches image gradients and $\mathcal{L}_{\mathrm{sep}}$ discourages lobe-direction collapse. Geometric guidance enters through bases and feature conditioning, without base-matching losses or cross-view parameter-consistency losses.

4 Experiments

Datasets. For public-benchmark experiments, all models are trained exclusively on the DL3DV-10K training set Ling et al. (2024). Following the DA3 evaluation protocol, we evaluate on the official 140-scene DL3DV-Benchmark, all Tanks and Temples training scenes except *Courthouse* Knapitsch et al. (2017), and Mip-NeRF 360 Barron et al. (2022).Models that support cam-

Backbone	Original (100k)			+PBE (50k)			+PBE (100k)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DL3DV									
AnySplat	21.73	0.7462	0.2920	22.38	0.7729	0.2906	22.64	0.7791	0.2942
YoNoSplat	23.96	0.8107	0.2156	25.05	0.8151	0.2616	25.22	0.8239	0.2493
VGGT-Ω + GSDPT	20.06	0.6829	0.3975	21.36	0.7725	0.3474	22.33	0.7384	0.3971
DA3	23.64	0.8267	0.2269	24.96	0.8747	0.1930	25.20	0.8794	0.1889
Mip-NeRF 360									
AnySplat	23.51	0.7448	0.2899	24.28	0.7771	0.2763	24.57	0.7826	0.2823
YoNoSplat	24.50	0.7612	0.2379	25.29	0.7485	0.2808	25.36	0.7516	0.2780
VGGT-Ω + GSDPT	21.92	0.6866	0.3856	23.47	0.7699	0.3331	24.54	0.7876	0.3299
DA3	24.55	0.7832	0.2668	26.01	0.8508	0.2220	26.42	0.8558	0.2187
Tanks and Temples									
AnySplat	20.49	0.7278	0.3355	21.02	0.7432	0.3283	21.27	0.7468	0.3313
YoNoSplat	22.05	0.7752	0.2621	23.51	0.7880	0.2793	23.60	0.7975	0.2703
VGGT-Ω + GSDPT	18.88	0.6548	0.4174	20.00	0.7246	0.3814	21.71	0.7265	0.3822
DA3	22.04	0.7759	0.2850	22.87	0.8151	0.2574	23.19	0.8228	0.2492

Table 1: Quantitative comparison on DL3DV, Mip-NeRF 360, and Tanks and Temples. Original denotes the baseline model, and +PBE denotes the model augmented with our module. Training budgets are reported in iterations. For each backbone and dataset, the best and second-best results for each metric across the three configurations are highlighted with light red and light blue backgrounds.

era estimation take unposed images as input and predict camera intrinsics and extrinsics internally. Other models use the camera parameters provided by the corresponding datasets.

Implementation details. We initialize the backbone and Gaussian head from official pretrained weights where available. When pretrained Gaussian-head weights are unavailable, as for VGGT-Ω, we attach a DA3-style Gaussian head. Optimization uses AdamW with a cosine learning-rate schedule and learning rates of 1×10^{-5} for the Gaussian head and 1×10^{-4} for PBE.For YoNoSplat, however, we set the learning rate to 1×10^{-6} for the Gaussian head and 1×10^{-5} for the PBE branch. All models are implemented in PyTorch and trained with mixed precision, FlashAttention, and a batch size of 1. For each architecture, the baseline and PBE share the same base-model checkpoint, training data, objectives, and optimization settings for their common components. Both variants are fine-tuned for 100K iterations for comparison under equal iteration budgets. We additionally evaluate PBE at 50K iterations to assess reconstruction quality with half the fine-tuning iterations.

All experiments are conducted on NVIDIA A5000 Pro GPU and NVIDIA A6000 GPU. For quantitative evaluation, we report PSNR, SSIM, and LPIPS to measure pixel-level fidelity, structural similarity, and perceptual quality, respectively. For each model, the baseline and its PBE variant are trained on the same GPU.

4.1 Comparisons

We evaluate PBE on dedicated feed-forward 3DGS methods, including AnySplat and YoNoSplat, and large-scale reconstruction models, including VGGT-Ω and DA3. We further evaluate cross-dataset generalization by training on DL3DV and testing on DL3DV-Benchmark, Tanks and Temples, and Mip-NeRF 360 using 16 input views, comparing each framework with and without PBE.

Table 1 shows consistent PSNR improvements across all four architectures and three benchmarks. PBE trained for 50K iterations outperforms the 100K-iteration baselines by 0.53–1.55 dB, while gains reach 0.78–2.83 dB at the same 100K budget. Notably, VGGT-Ω + GSDPT gains 2.83 dB on Tanks and Temples at 100K iterations, and DA3 gains 1.87 dB on Mip-NeRF 360 with only 100K iterations. SSIM and LPIPS improve in most settings, although perceptual gains are not uniform across models. Figure 3 compares light decomposition on the Lego scene across four baselines.

Backbone features	Depth features	Random priors	Physical priors	PSNR↑	SSIM↑	LPIPS↓
DL3DV						
✓				24.62	0.8680	0.2006
✓	✓			24.47	0.8638	0.2077
✓	✓	✓		24.73	0.8698	0.2024
✓	✓		✓	24.96	0.8747	0.1930

Table 2: Ablation of feature configurations and physical priors on DL3DV-Benchmark using 16 input views. The gray row denotes the full PBE model. The best and second-best results are highlighted with light red and light blue backgrounds, respectively.

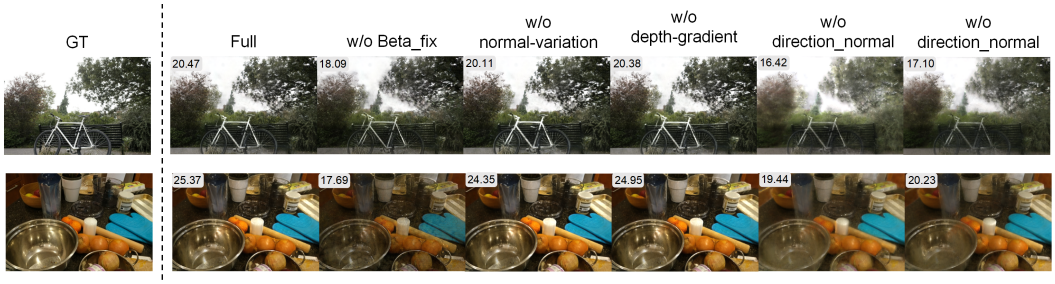


Figure 4: **Visualization of the ablation study.** We provide visualizations of the geometric-cue ablation on two scenes from the Mip-NeRF 360.

Compared with the baseline methods, our PBE module significantly improves the ability to capture specular components and can effectively separate the yellow highlights in the Lego scene. These results support the applicability of PBE across different reconstruction architectures and its ability to improve rendering fidelity with fewer fine-tuning iterations. They also show that PBE can effectively separate specular highlight components, thereby improving the rendering of specular reflections.

4.2 Ablation Study

Feature ablation. Using DA3, we compare backbone features only, features with depth, features with depth and random Beta bases, and full PBE. The first three feature variants retain the Beta/SB prediction architecture but remove explicit geometric guidance. All variants are trained on DL3DV for five epochs under the same strategy and evaluated on DL3DV-Benchmark with 16 input views. As shown in Table 2, PBE achieves the best performance across all three metrics, improving PSNR by 0.34 dB over features alone and 0.23 dB over random bases. Adding depth features alone does not improve performance, while random bases improve PSNR and SSIM but worsen LPIPS relative to features alone. These results support the value of explicitly constructed geometric bases beyond directly incorporating depth features or random bases.

Physical-prior ablation. We evaluate inference-time interventions on a fixed epoch-10 checkpoint using 16 input views on Mip-NeRF 360 and Tanks and Temples. Single disables multi-view base fusion while retaining multi-view evidence features, whereas Full retains fusion. All interventions modify single-view base candidates while keeping network weights and residual prediction paths unchanged; direction interventions additionally retain the original base colors and sharpness. As shown in Table 3, replacing the spatial Beta base with $\beta = 4$ under Full reduces PSNR from 26.44 to 20.04 dB on Mip-NeRF 360 and from 23.44 to 18.05 dB on Tanks and Temples. Removing the depth-gradient or normal-variation term produces smaller PSNR drops of 0.24–0.56 dB across the two datasets. Direction interventions cause substantial degradation: replacing the normal-based axis with the camera-to-surface ray reduces PSNR by 7.86 and 7.05 dB, respectively, while replacing the view-conditioned tangent frame with a camera-axis frame reduces PSNR by 6.07 and 5.25 dB. All interventions also worsen SSIM and LPIPS, with similar trends under Single. These input-view results show that the trained residual predictor does not fully compensate for changes to the

Variant	Mip-NeRF 360 PSNR↑	Mip-NeRF 360 SSIM↑	Mip-NeRF 360 LPIPS↓	Tanks and Temples PSNR↑	Tanks and Temples SSIM↑	Tanks and Temples LPIPS↓
Single-view bases (Single)						
Constant β base (4)	19.94	0.7753	0.3242	18.05	0.7647	0.3167
w/o depth-gradient term	25.86	0.8535	0.2198	23.05	0.8224	0.2504
w/o normal-variation term	25.74	0.8502	0.2198	23.19	0.8227	0.2499
w/o direction normal	18.56	0.6220	0.4025	16.37	0.6304	0.4184
w/o direction view tangent	20.45	0.6829	0.3684	18.20	0.6990	0.3661
Original geometric bases (Ours)	26.45	0.8608	0.2120	23.45	0.8309	0.2394
Multi-view base fusion (Full)						
Constant β base (4)	20.04	0.7791	0.3198	18.05	0.7652	0.3161
w/o depth-gradient term	25.90	0.8547	0.2192	23.05	0.8228	0.2499
w/o normal-variation term	25.88	0.8526	0.2182	23.20	0.8233	0.2493
w/o direction normal	18.59	0.6234	0.4024	16.38	0.6308	0.4183
w/o direction view tangent	20.37	0.6832	0.3689	18.18	0.6991	0.3663
Original geometric bases (Ours)	26.44	0.8616	0.2117	23.44	0.8312	0.2391

Table 3: **Ablation of geometric bases.** All variants share the epoch-10 checkpoint and use 16 views per scene. Single disables multi-view base fusion but keeps multi-view evidence features; Full retains fusion. Interventions modify single-view bases before fusion. Direction variants keep original base colors and sharpness.

geometric bases; they characterize model dependence at inference rather than the performance of alternative designs after retraining. The qualitative comparisons in Figure 4 further demonstrate the effectiveness of physical priors, which better preserve fine textures and object boundaries.

These ablations show that PBE's explicit use of geometric information is effective and superior to alternative designs. For the original bases, Single yields slightly higher PSNR, whereas Full yields slightly higher SSIM and lower LPIPS on both datasets.

5 Conclusion

This paper presents PBE, a lightweight extension module for feed-forward 3DGS. We demonstrate that, without modifying or retraining the pretrained backbone, introducing a module with approximately 0.35M parameters is sufficient to improve Gaussian prediction across diverse architectures. PBE constructs explicit Beta/SB parameter bases from geometric cues and support-view observations, and refines them using learned features. Extensive experiments show consistent improvements across multiple models and public benchmarks, especially in specular reflection scenes, while reducing training time by 41–48 % to reach comparable performance.

AI use statement

Generative AI was used solely for language polishing, including improving the grammar, fluency, and readability of the manuscript. It did not contribute to the research ideas, methodology, experiments, data analysis, or technical content. We have reviewed all AI-assisted content and take responsibility for the final content of this work, including text, claims, or artifacts polished with the aid of generative AI tools. All polishing was limited to language expression and did not alter the research design, data, results, conclusions, or academic claims of the paper.

Reproducibility statement

We provide the complete formulation of PBE and its training method in the Method section, and present the training data, evaluation protocols, and implementation details in experiments. The detailed computation of PBE is further described in the appendix.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469. IEEE, 2022.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467. IEEE, 2024.

Stavros Diolatzis, Tobias Zirr, Alexander Kuznetsov, Georgios Kopanas, and Anton Kaplanyan. N-dimensional gaussians for fitting of high dimensional functions. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Zhongpai Gao, Benjamin Planche, Meng Zheng, Anwesa Choudhuri, Terrence Chen, and Ziyan Wu. 6dgs: Enhanced direction-aware gaussian splatting for volumetric rendering. In *International Conference on Learning Representations*, volume 2025, pp. 30884–30902, 2025.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Rong Liu, Dylan Sun, Meida Chen, Yue Wang, and Andrew Feng. Deformable beta splatting. In *Proceedings of the special interest group on computer graphics and interactive techniques conference conference papers*, pp. 1–11, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω, 2026a. URL https://arxiv.org/abs/2605.15195.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16453–16463. IEEE, 2025.

Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20923–20931. IEEE, 2024.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 2025.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. Yonosplat: You only need one model for feedforward 3d gaussian splatting. In *International Conference on Learning Representations*, volume 2026, pp. 39852–39871, 2026.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19447–19456. IEEE, 2024.

Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. *Advances in Neural Information Processing Systems*, 37:77799–77830, 2024.