Supplementary Material

This Supplementary Material provides additional technical details, ablation studies, and qualitative evaluations that complement the main paper. Section A reports the full implementation details and hyperparameters for all modules in DIAMOND-SSS. Section B provides additional qualitative reconstructions under a wide range of capture regimes, demonstrating the robustness of DIAMOND-SSS, particularly in sparse-view and sparse-light settings where SSS-3DGS fails severely. Section C details the fine-tuning procedures for the diffusion-based components—multi-view and relighting diffusion—and presents extended qualitative comparisons against off-the-shelf NVS and relighting baselines. Section D expands on the formulation and practical considerations of the multi-view geometric losses, including visibility handling, depth normalization, sampling strategy, and targeted ablation studies on both synthetic and real objects. Finally, Section E summarizes compute requirements, reproducibility notes, and implementation considerations to facilitate future research and adoption.

A. Implementation and Hyperparameters

In this appendix we provide implementation and optimization details that complement the main paper (see Sec. 4.2).

A.1. Optimization and Training Schedule

We jointly optimize the 3D Gaussian parameters and the SSS residual MLP (see Sec. 3.1) using the Adam optimizer, with separate learning rates for geometry (5×10^{-3}) and appearance (1×10^{-3}). Both learning rates are reduced by a factor of 0.1 after 50k iterations. Training runs for 80k–100k steps depending on the object and capture regime.

Gaussian initialization, densification, and pruning follow the SSS-3DGS pipeline [7]. To ensure fair comparison across ablations, we keep the number of Gaussians fixed throughout optimization.

All experiments are implemented in PyTorch 2.2 with mixed precision and executed either on a single NVIDIA RTX A6000 (48 GB) or on a workstation equipped with a GPU with at least 24 GB VRAM, 32 GB RAM, and roughly 100 GB of disk space. The environment uses CUDA 11.6 for compatibility with our training framework.

A.2. Loss Weights

The main paper defines the full objective $\mathcal{L}_{\text{total}}$ in Sec. 3.3, combining photometric, regularization, and geometric consistency terms. For completeness, we summarize the scalar weights used in all experiments.

We employ a mixed photometric objective

$$\begin{aligned}\mathcal{L}_{\text{total}} =&(1-\lambda_{\text{dssim}})\mathcal{L}_1 + \lambda_{\text{dssim}}(1-\text{SSIM}) + \lambda_{\text{lpips}}\mathcal{L}_{\text{LPIPS}} \\ &+ \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{enh}}\mathcal{L}_{\text{enh}} \\ &+ \lambda_{\text{ray}}\mathcal{L}_{\text{ray}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}}^{\text{MV}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}^{\text{MV}},\end{aligned} \tag{A.1}$$

The regularization objective proposed in [7] combines several auxiliary losses that improve stability, material fidelity, and geometric consistency. First, a normal-consistency term enforces agreement between the predicted normals $\mathbf{N}$ and pseudo-normals $\tilde{\mathbf{N}}$ derived from the rendered depth map D under a local planarity assumption:

$$\mathcal{L}_{\text{normals}} = \lambda_{\text{normals}}\|\mathbf{N} - \tilde{\mathbf{N}}\|_2. \tag{A.2}$$

To supervise incident radiance, we constrain the clamped predicted illumination $\overline{L}_{\text{in}}(x, \boldsymbol{\omega_i})$ to match the learned spherical-harmonics visibility $V_{SH}(x, \boldsymbol{\omega_i})$ via an L1 loss

$$\mathcal{L}_{\text{incident}} = \lambda_{\text{incident}}\left\| \overline{L}_{\text{in}}(x, \boldsymbol{\omega_i}) - V_{SH}(x, \boldsymbol{\omega_i})\right\|_1. \tag{A.3}$$

Foreground consistency is enforced with a spatial masking loss, penalizing the contribution of Gaussians outside the image mask I_{mask}:

$$\mathcal{L}_{\text{mask}} = -\lambda_{\text{mask}}\left[I_{\text{mask}}\log(\alpha) + (1-I_{\text{mask}})\log(1-\alpha)\right]. \tag{A.4}$$

Material smoothness is encouraged using bilateral smoothing on metalness m, roughness r, subsurfaceness sss, and base color $\mathbf{b}$, each with its own weight $\lambda_{\text{smooth}}^{(q)}$; for a generic attribute PI_q the loss is

$$\mathcal{L}_{\text{smooth}}^{(q)} = \frac{1}{N}\sum_{x,y} M(x,y)\,\|\nabla PI_q(x,y)\|\,\exp\left(-\|\nabla I(x,y)\|\right), \tag{A.5}$$

where $I(x,y)$ is the input image and $M(x,y)$ denotes the valid mask. To improve highlight and shadow reconstruction, the authors in [7] adopt the R3DGS enhancement loss [9], comparing the predicted base color $\mathbf{b}$ to a pseudo-target T constructed from the input RGB image I_{rgb} using a sigmoid-based blending weight sw:

$$\begin{aligned}\mathcal{L}_{\text{enhance}} &= \lambda_{\text{enhance}}\|T - \mathbf{b}\|_1, \\ T &= sw \cdot I_{rgb}^2 + (1-sw)\cdot\left[1-(1-I_{rgb})^2\right].\end{aligned} \tag{A.6}$$

Finally, we supervise learned visibility using ray-traced ground-truth visibility V_{RT}:

$$\mathcal{L}_{\text{raytrace}} = \lambda_{\text{raytrace}}\left\|V_{SH}(x, \boldsymbol{\omega_i}) - V_{RT}(x, \boldsymbol{\omega_i})\right\|_1. \tag{A.7}$$

Together, these terms form the complete regularization objective used during optimization. All hyperparameters used in our experiments are presented in Tab. 3.

A.3. Hyperparameter Tuning

The goal in this section is to understand how different configurations affect the custom objective $\mathcal{L}_{\text{total}}$ (see Eq. A.1), which aggregates photometric and regularization terms into a single performance metric for translucent scenes.

A.3.1. Parallel Coordinate Analysis

Figure 5 shows a parallel coordinate plot where each polyline corresponds to one hyperparameter configuration, and each vertical axis represents a different hyperparameter. The color encodes performance: lighter lines denote higher $\mathcal{L}_{\text{total}}$ (better configurations), while darker lines correspond to lower-performing settings.

Several trends emerge:

- Mid-range values of λ_{dssim} (0.25,0.3) and $\lambda_{\text{incident\_light}}$ (0.01-0.012) tend to perform best, suggesting that moderate weighting of structural similarity and incident-light regularization provides a good trade-off for reconstruction quality.
- Higher values of λ_{normal} (0.045>) and $\lambda_{\text{base\_color\_smooth}}$ (0.018>) correlate with better performance, indicating that emphasizing surface consistency and base-color smoothness can enhance overall fidelity.
- For $\lambda_{\text{roughness\_smooth}}$ (0.002<) and `densify_until_iter` (10K<), the best configurations cluster towards lower values, suggesting that overly strong roughness smoothing or excessively long densification phases can harm fine detail and lead to suboptimal reconstructions.

Overall, these patterns provide guidance on how to balance detail preservation, smoothness, and physical plausibility in SSS-3DGS-style reconstructions.

A.3.2. Best Configuration vs. Original

Table 4 compares the original hyperparameter configuration from [7] with the best-performing combination identified by our tuning procedure.

A few observations:

- The perceptual loss weight λ_{lpips} decreases from 0.2 to 0.171, suggesting that a slightly reduced emphasis on perceptual similarity can improve the aggregate objective when combined with other terms.
- In contrast, the structural loss weight λ_{dssim} increases substantially from 0.2 to 0.492, indicating that stronger structural supervision plays a key role in achieving better reconstructions.
- The normal-consistency term λ_{normal} is tuned upward from 0.02 to 0.037, reinforcing the benefit of encouraging local surface coherence.
- Among regularization terms, $\lambda_{\text{metallic\_smooth}}$ and $\lambda_{\text{roughness\_smooth}}$ increase, whereas $\lambda_{\text{base\_color}}$ and $\lambda_{\text{base\_color\_smooth}}$ are slightly reduced. This suggests that more flexible modeling of specular behavior combined with milder regularization on base color yields better fidelity.
- The densification strategy favors a shorter and more aggressive phase: `densify_until_iter` is reduced from 15000 to 10000, `densification_interval` is slightly increased, and `densify_grad_threshold` is relaxed. The opacity reset interval is also reduced from 3000 to 1000. Together, these changes promote faster adaptation of the Gaussian set and help avoid early overfitting.

Overall, the tuned configuration reflects a nuanced rebalancing of perceptual, structural, and regularization terms, leading to improved performance under the aggregated objective $\mathcal{L}_{\text{total}}$ while remaining close in spirit to the original SSS-3DGS design.

B. Additional Qualitative Reconstructions

To complement the quantitative evaluations presented in the main paper, we provide additional qualitative reconstructions across a variety of translucent objects and capture regimes. These visual comparisons illustrate how DIAMOND-SSS behaves under different levels of supervision, ranging from full-view OLAT captures to extremely sparse pose–light subsets.

Comparison across objects and capture regimes. Figure 6 presents reconstructions for three representative objects—*Plastic Bottle*, *Crystal*, and *Massage Ball*—under four supervision levels: (i) all views and all lights, (ii) all views with a single light per view, (iii) 5% of views and 5% of lights, and (iv) 3% of views and 3% of lights.

In all settings, DIAMOND-SSS produces sharper boundaries, more stable silhouettes, and more faithful subsurface-scattering cues than SSS-3DGS [7]. In the sparse regimes (5%/5% and 3%/3%), SSS-3DGS fails severely—exhibiting strong boundary leakage, collapsed geometry, and inconsistent opacity—while DIAMOND-SSS remains stable and reconstructs the correct translucent appearance.

Fine-grained appearance comparisons. Figure 7 provides additional close-up views highlighting translucent appearance, color diffusion, and internal glow. Across objects, DIAMOND-SSS preserves the soft radiance falloff characteristic of translucent materials while maintaining geometric fidelity. In contrast, SSS-3DGS often exhibits noisy silhouettes, incorrect color bleeding near boundaries, or inconsistent opacity when supervision becomes sparse.

These qualitative results reinforce the core claim of the paper: by integrating diffusion-based augmentation with multi-view geometric consistency, DIAMOND-SSS remains

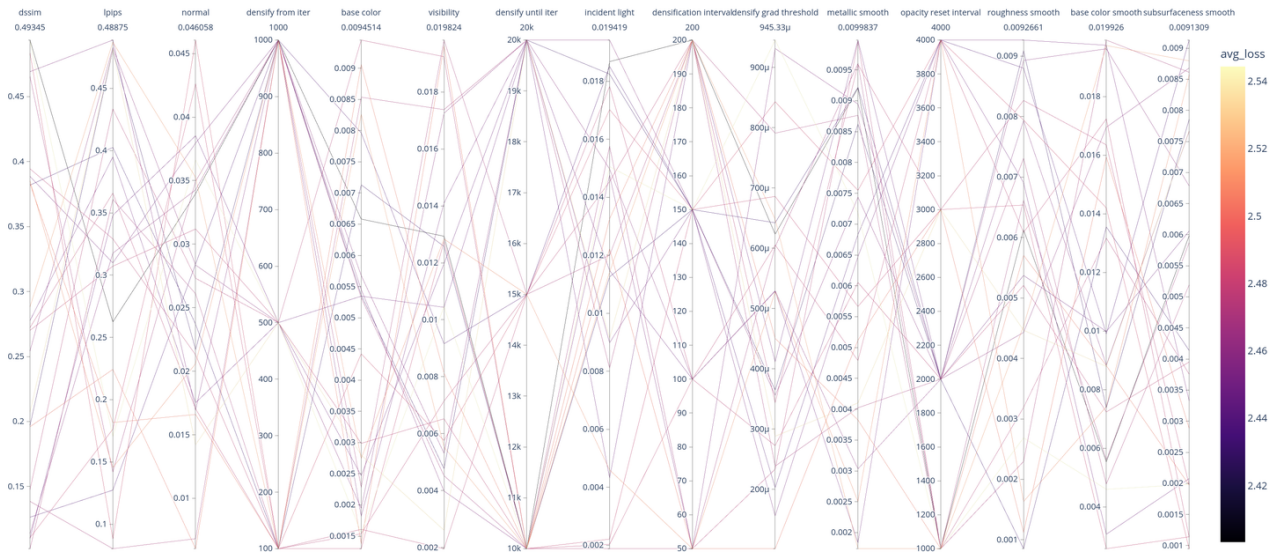


Figure 5. **Parallel coordinate plot of the hyperparameter search.** Each line corresponds to one configuration; lighter colors indicate higher values of the aggregated objective $\mathcal{L}_{\text{total}}$, as defined in Eq. A.1.

Training and learning rate parameters		**Densification and loss parameters**	
Configuration	**Value** [7]	**Configuration**	**SValue**
iterations	10,000	percent_dense	0.01
batch size	4	densification_interval	150
Gaussian learning rate	5×10^{-3}	opacity_reset_interval	3000
MLP learning rate	1×10^{-3}	densify_from_iter	500
position_lr_init	0.00016	densify_until_iter	10000
position_lr_final	0.0000016	densify_grad_threshold	0.0010
position_lr_delay_mult	0.01	densify_grad_normal_threshold	–
position_lr_max_steps	30,000	normal_densify_from_iter	0
normal_lr	0.01	random_background	False
normals_lr	0.01	sss_width	32
sh_lr	–	λ_{dssim}	0.492
feature_lr	0.0025	λ_{lpips}	0.171
color_lr	0.0025	λ_{mask}	0.1
opacity_lr	0.05	λ_{smooth}	[0.002, 0.002, 0.002, 0.006]
scaling_lr	0.005	λ_{enh}	0.005
rotation_lr	0.001	λ_{ray}	0.01
env_lr	0.1	λ_{sil}	0.5
env_rest_lr	0.001	λ_{depth}	0.3
base_color_lr	0.01	α (synthetic weight)	0.5
roughness_lr	0.01	λ_{normal}	0.037
metallic_lr	0.01	$\lambda_{\text{visibility}}$	0.01
subsurfaceness_lr	0.01	$\lambda_{\text{incident\_light}}$	0.015
light_lr	0.001	$\lambda_{\text{mask\_entropy}}$	0.1
light_rest_lr	0.0001	$\lambda_{\text{base\_color}}$	0.002
light_init	-1.0	$\lambda_{\text{base\_color\_smooth}}$	0.005
visibility_lr	0.0025	$\lambda_{\text{metallic\_smooth}}$	0.007
visibility_rest_lr	0.0025	$\lambda_{\text{roughness\_smooth}}$	0.003
sss_lr	0.001	$\lambda_{\text{subsurfaceness\_smooth}}$	0.002

Table 3. **DIAMOND-SSS configuration** Training and densification hyperparameters for DIAMOND-SSS.

robust even under highly incomplete OLAT captures, providing state-of-the-art reconstructions of translucent materials.

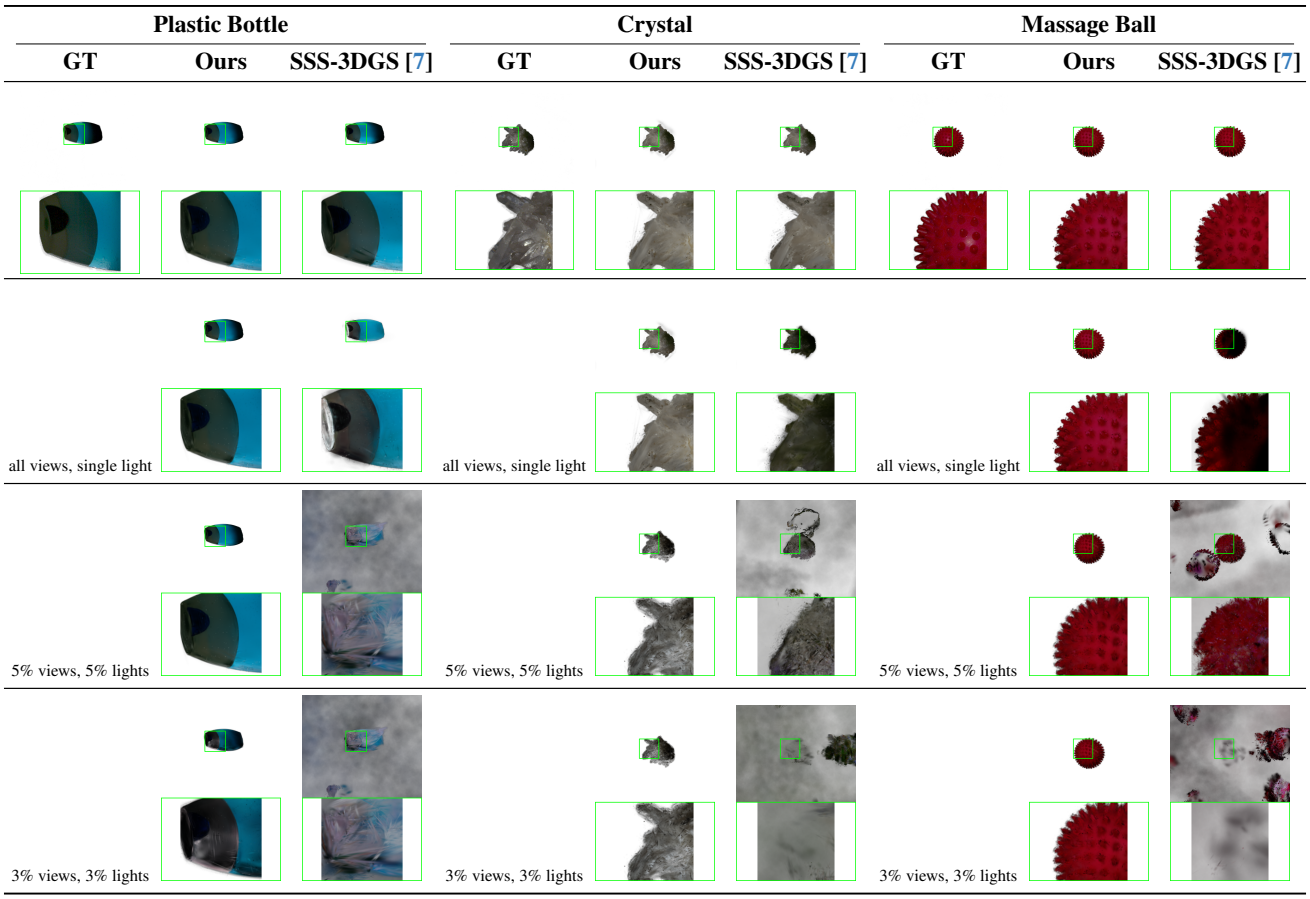


Figure 6. Qualitative comparison of reconstructed translucent appearance under different supervision conditions, the first row shows the setting: all views, all lights.

Hyperparameter	Original	Best
λ_{dssim}	0.200	0.492
λ_{lpips}	0.200	0.171
λ_{normal}	0.020	0.037
$\lambda_{\text{visibility}}$	0.010	0.010
$\lambda_{\text{incident\_light}}$	0.020	0.015
$\lambda_{\text{base\_color}}$	0.005	0.002
$\lambda_{\text{base\_color\_smooth}}$	0.006	0.005
$\lambda_{\text{metallic\_smooth}}$	0.002	0.007
$\lambda_{\text{roughness\_smooth}}$	0.002	0.003
$\lambda_{\text{subsurfaceness\_smooth}}$	0.002	0.002
densification_interval	100	150
opacity_reset_interval	3000	1000
densify_from_iter	500	500
densify_until_iter	15000	10000
densify_grad_threshold	0.0002	0.0010

Table 4. **Original vs. tuned hyperparameter configuration.** The "Best" setting corresponds to the highest $\mathcal{L}_{\text{total}}$ found in our search, as illustrated in Fig. 5.

C. Diffusion Fine-Tuning Details

The diffusion-based augmentation pipeline is introduced in Sec. 3.2 and integrated into the reconstruction procedure in Sec. 3.4. Here we provide additional training and conditioning details for the two diffusion models used in DIAMOND-SSS: a multi-view (novel-view) diffusion model and a relighting diffusion model. Both models are fine-tuned once on a small OLAT subset and are then reused across all test objects and capture regimes, without any per-object retraining.

C.1. Training Data for Diffusion Models

Figure 8 shows an overview of the dataset and Fig. 9 depicts the OLAT capture setup used by [7] to generate the dataset. As described in Secs. 4.1 and 4.2, both diffusion models are fine-tuned on a small subset ($\leq 7\%$) of the full OLAT training data from SSS-3DGS [7]. We select four representative translucent objects—*red car*, *jam jar*, *wax candle*, and *marble head*—covering a range of scattering behaviors and specularities.

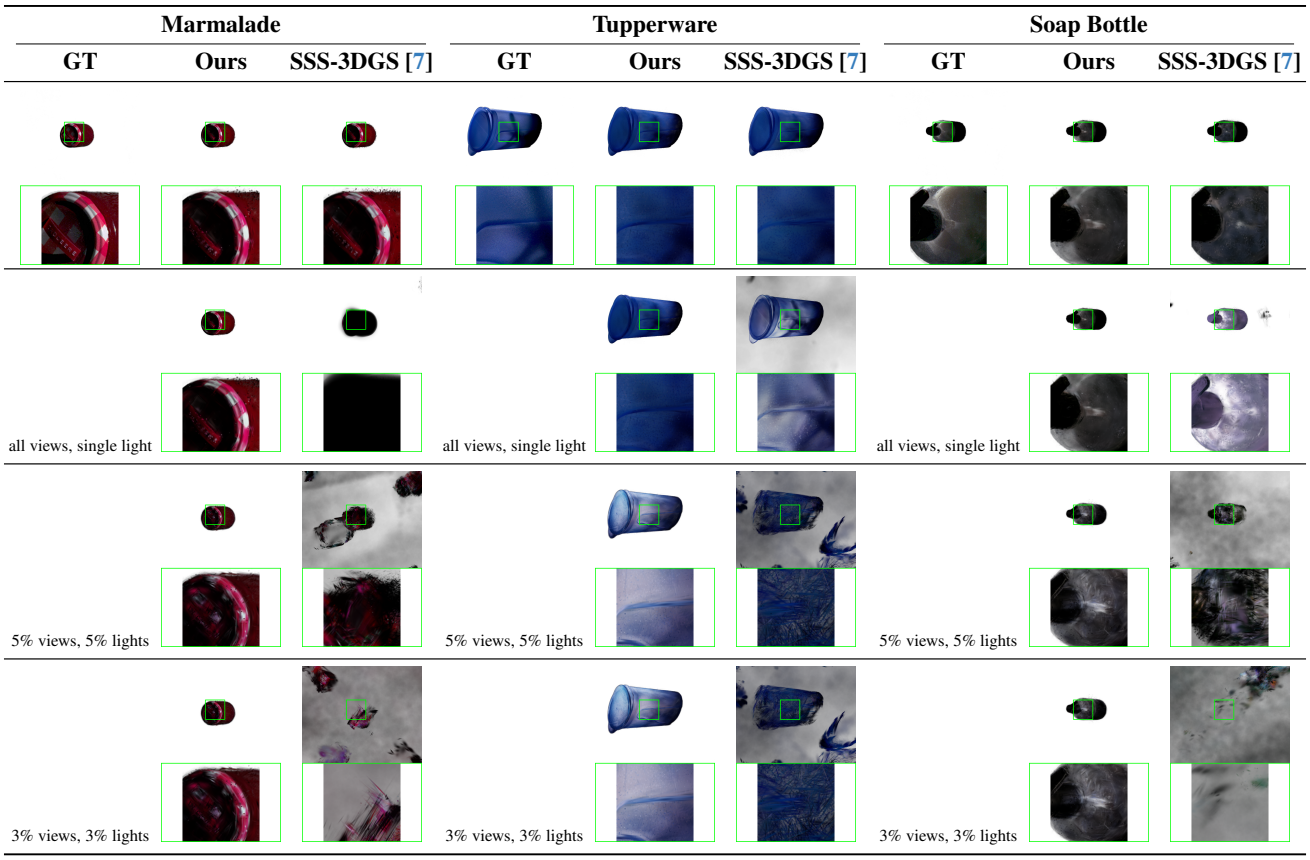


Figure 7. Qualitative comparison of reconstructed translucent appearance under different supervision conditions, the first row shows the setting: all views, all lights.

We emphasize that:

- Fine-tuning uses only training views and illuminations; no evaluation frames are seen by the diffusion models.
- The same fine-tuned models are reused across *all* test objects and capture regimes (no per-object or per-regime retraining).
- Illumination is kept fixed for multi-view diffusion (to isolate geometry) and view is kept fixed for relighting diffusion (to isolate light transport).

Together, these two diffusion models are used to expand a sparse OLAT capture into a denser multi-view, multi-illumination training set for DIAMOND-SSS.

C.2. Multi-View Diffusion (Novel-View Synthesis)

Purpose and model selection. Given a reference image I_{v_r} from viewpoint v_r and a target pose $P_v = (K, T_v)$, the multi-view diffusion module synthesizes the corresponding novel view $\hat{I}_v$. This densifies the camera trajectory under fixed illumination, enabling cross-combination with the relighting module to obtain synthetic multi-view, multi-light supervision within DIAMOND-SSS.

We first benchmarked several diffusion-based NVS models in inference mode—Zero123 [19], Zero123++ [27], MVD-Fusion [11], and Free3D [45]. On SSS-3DGS OLAT data, Free3D showed the strongest geometry and silhouette consistency out of the box, and we adopt it as our baseline prior. Representative comparisons are provided in Figs. 10 to 12. See Fig. 2 for an overview of the architecture.

Training procedure. We fine-tune the official Free3D checkpoint [45] (latent-diffusion UNet, 32×32 latents, scaling 0.18215) using the original denoising objective and linear noise schedule ($0.00085 \rightarrow 0.0120$ over 1000 steps). Training sequences are sampled under fixed illumination so that the model learns geometric variation without entangling lighting.

Camera poses are injected through ray-conditioning normalization layers, and lightweight multi-view attention and noise-sharing modules promote cross-view consistency. No architectural changes are introduced; only the weights are adapted to the OLAT domain.

Fine-tuning hyperparameters:

- **Batch size:** 16 target views sampled from 4 distinct objects,

- **Learning rate:** 1×10^{-4} with 100 warm-up steps,
- **Training duration:** ≈ 500 epochs on the OLAT subset,
- **Memory optimization:** Gradient checkpointing enabled.

Use within DIAMOND-SSS. During reconstruction, synthetic views generated by the multi-view diffusion module are rendered through the same camera poses as real observations. Photometric losses on synthetic images are down-weighted by the factor α defined in Sec. 3.4, while geometric consistency losses (e.g., silhouette or depth consistency) benefit fully from the increased viewpoint density.

C.3. Relighting Diffusion

Purpose and model selection. Given a single-view RGB input I_{v,l_s} under source illumination l_s, together with auxiliary conditioning maps C_v (depth and normals), the relighting diffusion module predicts the same view under a novel target light l_{tgt}. This expands the illumination coverage under fixed geometry and is especially useful in regimes such as "all views, 1 light per view," where synthetic OLAT variants provide dense lighting supervision.

We adopt a ControlNet-style conditional diffusion framework inspired by [23, 40], starting from the public implementation of Poirier-Ginter *et al*. [23]. An architectural overview is included in Fig. 2.

Training procedure. We augment the model original model to receive the following conditioning signals:
- **Source RGB image:** I_{v,l_s},
- **Estimated depth map:** from Depth Anything v2 [35],
- **Estimated surface normals:** from Marigold [16],
- **Target illumination:** 9D spherical harmonic encoding $\Phi(l_{tgt})$.

Depth and normals are resized and normalized to match the UNet resolution. SH features are injected into both the ControlNet branch and the denoiser's timestep and cross-attention embeddings. The architecture itself is unchanged; only the training data and conditioning inputs are adapted to OLAT captures. In Fig. 13 depicts examples of the conditions used per object.

Training objective. We employ the v-parameterization objective combined with a mixture of L1, SSIM, and LPIPS losses, as well as cycle-consistency and blur-aware terms to encourage stable low-frequency light transport and reduce high-frequency artifacts. Training samples follow:

$$(I_{v,l_s},\, D_v,\, N_v,\, \Phi(l_{tgt}),\, I_{v,l_{tgt}}),$$

where $I_{v,l_{tgt}}$ is the corresponding OLAT ground truth. The implementation follows [23], with modifications only to the data and conditioning.

Use within DIAMOND-SSS. During reconstruction, the relighting diffusion module produces synthetic OLAT variants for each view. As with multi-view diffusion outputs, synthetic photometric losses are down-weighted by α, but these augmented illuminations still contribute meaningfully to multi-view silhouette and depth consistency, improving supervision across illumination conditions.

C.4. Setup: Use with Real, Relighted, and Fully Synthetic Data

Real OLAT captures. We use native OLAT images with provided (K_v, T_v) without modification.

Relighting augmentation. For settings such as "all views, 1 light per view," missing illuminations are synthesized using the relighting diffusion model (Sec. C.3). The camera parameters (K_v, T_v) remain unchanged.

Fully synthetic multi-view sets. When using multi-view diffusion + relighting, we generate:
- novel poses P_v from the Free3D-based NVS model (Sec. C.2),
- relighted OLAT equivalents under all target directions.

Generated data is aligned with the original OLAT lattice (same lights, comparable poses).

C.5. Benchmarking and ablation studies

We evaluate the effect of diffusion fine-tuning for both the multi-view synthesis module and the relighting module. Figures 10–12 summarize the multi-view synthesis behavior across several baselines, while Figures 14 and 15 focus on the relighting model. In addition, Table 2 shows our numerical results for the ablation studies.

Multi-view diffusion (NVS). We compare our fine-tuned Free3D model with: (1) off-the-shelf Free3D, (2) Zero123, (3) Zero123-XL, and (4) MVD-Fusion.

Across synthetic and real objects (Figures 10, 11, 12), our fine-tuned model consistently produces:
- sharper and more stable silhouettes,
- reduced geometric drift across viewpoints,
- fewer distortions in thin structures,
- better preservation of object identity under large viewpoint changes.

This justifies using Free3D as the backbone for DIAMOND-SSS and highlights the importance of domain adaptation to OLAT capture conditions.

Relighting diffusion. We further ablate the effects of progressively increasing conditioning and supervision:
1. Baseline RGB-only conditioning,
2. + depth and normal conditioning,

3. + L1 and perceptual losses.

As illustrated in Figures 14 and 15, each addition improves physical realism:

- depth improves global shadow placement,
- normals sharpen high-frequency surface cues,
- perceptual losses reduce haloing and over-sharp reflections,
- the full model best reproduces soft-scattering cues, achieving smooth light transitions characteristic of SSS material appearance.

Summary. The ablations show that diffusion models benefit substantially from fine-tuning under our OLAT dataset and conditioning strategy. For NVS, fine-tuning improves geometric coherence and silhouette stability. For relighting, deeper conditioning leads to more faithful subsurface-scattering behavior and reduced artifacts. These findings support the design choices adopted in DIAMOND-SSS and demonstrate that diffusion-based augmentation becomes reliable only after domain adaptation.

D. Multi-View Geometric Losses: Practical Details

The multi-view geometric consistency losses introduced in Sec. 3.3—namely multi-view silhouette consistency and multi-view depth consistency—play a central role in stabilizing reconstructions under sparse or synthetic supervision. This section provides additional implementation details, motivation, and qualitative analyses that complement the main text.

D.1. Motivation and Intuition

Multi-view photometric supervision alone does not sufficiently constrain the geometry of translucent objects, in particular in our very sparse setting with synthetic data augmentation. Small inconsistencies in predicted depth or occupancy can lead to view-dependent artifacts, such as boundary bleeding or inconsistent surface placement, which propagate into both shading and appearance, especially under SSS transport.

The two geometric consistency terms complement each other:

Silhouette Consistency. The silhouette-consistency term enforces that pixels corresponding to the object's foreground in one view remain foreground when reprojected into another view. Concretely, for a pixel x in view i, we back-project it using its predicted depth and reproject it into view j (see Eq. (3.5)). The loss penalizes cases where this reprojected point falls outside the soft silhouette mask of view j, thereby stabilizing object boundaries, reducing opacity bleeding, and improving cross-view alignment of object outlines.

Depth Consistency. Silhouette alignment alone does not provide depth information. The depth consistency loss addresses this by enforcing agreement between the predicted metric depth in view j and the back-projected/reprojected depth from view i (see Eq. (3.6) in the main paper). This ensures that the same physical surface point occupies a consistent 3D position across camera viewpoints, thereby improving global geometric coherence.

Together, these losses constrain both *where* the surface should appear (silhouette) and *how far* it should lie along each viewing ray (depth), reducing cross-view inconsistencies and improving stability under downstream relighting.

D.2. Visibility and Correspondence Filtering

Both $\mathcal{L}_{\text{sil}}^{\text{MV}}$ and $\mathcal{L}_{\text{depth}}^{\text{MV}}$ rely on correspondences obtained via back-projection from a source view i to a target view j using the camera models. In practice, we improve robustness by:

- **Domain filtering:** Correspondences whose reprojected coordinates $x'_{i,j}$ fall outside the target image plane are discarded.
- **Silhouette filtering:** Pixels reprojected outside the target soft silhouette (beyond a small tolerance) are ignored.
- **Depth validity:** Correspondences with invalid or missing depth (e.g., occlusions or empty space) in either view are removed.
- **Normalization:** Each loss term is normalized by the number of valid correspondences to avoid biasing toward dense or front-facing views.

This visibility-aware masking substantially reduces artifacts near occlusion boundaries and improves stability in sparse-view settings.

D.3. Depth Normalization and Numerical Stability

Depth values depend on the scene scale and camera intrinsics. To improve numerical conditioning, both the reprojection depth $\hat{z}_{i,j}$ and the rendered depth $D_j(x')$ are linearly normalized to $[0, 1]$ using scene-specific near/far clipping planes. This removes the need for per-object tuning of λ_{depth} and yields consistent gradients across scenes.

D.4. Ablation Studies

We evaluate the contribution of each multi-view geometric term through controlled ablations on synthetic and real OLAT data. The quantitative results are summarized in Table 2, and qualitative comparisons appear in Figures 16–21.

Silhouette-only supervision. Adding only the silhouette consistency term improves boundary alignment by ensuring that reprojected points fall within the target-view silhouette.

As shown in Figure 16 (second column), this reduces edge bleeding and stabilizes contours but does not fully resolve metric inconsistencies.

Depth-only supervision. The depth consistency loss enforces agreement between the reprojected depth $\hat{z}_j(X)$ and the rendered depth $D_j(x')$ in the target view. Figure 16 (third column) shows that this significantly reduces geometric drift across views, improving spatial coherence even in sparse-view settings.

Combined silhouette + depth supervision. Using both terms yields the most stable reconstructions. As seen in Figures 16 (fourth column) and 17, the combination delivers accurate geometry and clean boundaries, with fewer view-dependent artifacts than either component alone.

Generalization across objects. Figures 18 and 19 demonstrate that these improvements generalize across diverse synthetic and real objects, including highly translucent materials. The combined loss consistently produces sharper silhouettes and more metrically consistent shapes.

Isolated silhouette effects. Figures 20 and 21 highlight the specific gain from silhouette consistency alone—most notably the suppression of boundary haloing and improved contour sharpness.

Summary. Across all experiments, multi-view geometric supervision (i) sharpens boundaries, (ii) improves metric alignment, and (iii) reduces view-dependent distortions. These trends match the improvements reported quantitatively in Table 2 and play a central role in the reliability of DIAMOND-SSS under sparse or noisy view regimes.

E. Reproducibility and Compute

DIAMOND-SSS is built entirely on publicly released components in 3D Gaussian Splatting and diffusion architectures. To facilitate reproducibility, we summarize below the practical considerations, compute settings, and implementation choices that complement the hyperparameters in Tab. 3 and the training details provided in Secs. C and 4.2.

Hardware. All experiments were conducted on a single NVIDIA RTX A6000 GPU (49 GB VRAM) using PyTorch 2.2 with automatic mixed precision. A local workstation with 32 GB RAM and ∼100 GB of disk storage (datasets, checkpoints, intermediate outputs) is sufficient to reproduce every experiment. Training the Free3D fine-tuning and the relighting model does not require multi-GPU setups.

Code base. Our implementation builds on:
- **3DGS** [17] for the base Gaussian renderer,
- **Relightable 3DGS** [9] for material and lighting branches,
- **SSS-3DGS** [7] for subsurface-scattering modeling,
- **Free3D** [45] for multi-view diffusion,
- **Poirier-Ginter et al.** [23] for relighting diffusion (ControlNet-style conditioning).

All components required to reproduce our results are open source.

Training schedule and runtime. Reconstruction of each scene (including multi-view geometric consistency) requires approximately:
- **1–1.5 hour** for **full-view** OLAT settings,
- **0.5–1 hours** for **sparse-view** settings,

measured on an RTX A6000. Fine-tuning the diffusion models is performed once:
- **Free3D NVS fine-tuning:** ≈15 hours for 500 epochs,
- **Relighting model:** ≈12–20 hours depending on object subset size.

These models are then reused for all experiments without per-object retraining.

Scene reuse and caching. Rendered depth, normals, and intermediate diffusion results are cached to reduce repeated compute. Gaussian densification schedules and multi-view loss masks (Sec. D) are shared across experiments to ensure comparability.

Randomness and determinism. All experiments use fixed seeds for camera sampling, appearance model initialization, and diffusion noise, ensuring repeatability with no run-to-run variations.

Reconstruction pipeline consistency. Evaluation scripts, loss computation, and image-space metrics (PSNR, SSIM, LPIPS) follow the same routines across synthetic and real scenes for consistent reporting.

Overall, DIAMOND-SSS can be reproduced on a single high-memory GPU and relies only on publicly available components. The training schedules, hyperparameters, and data augmentations are fully documented in the main paper and this supplementary material.

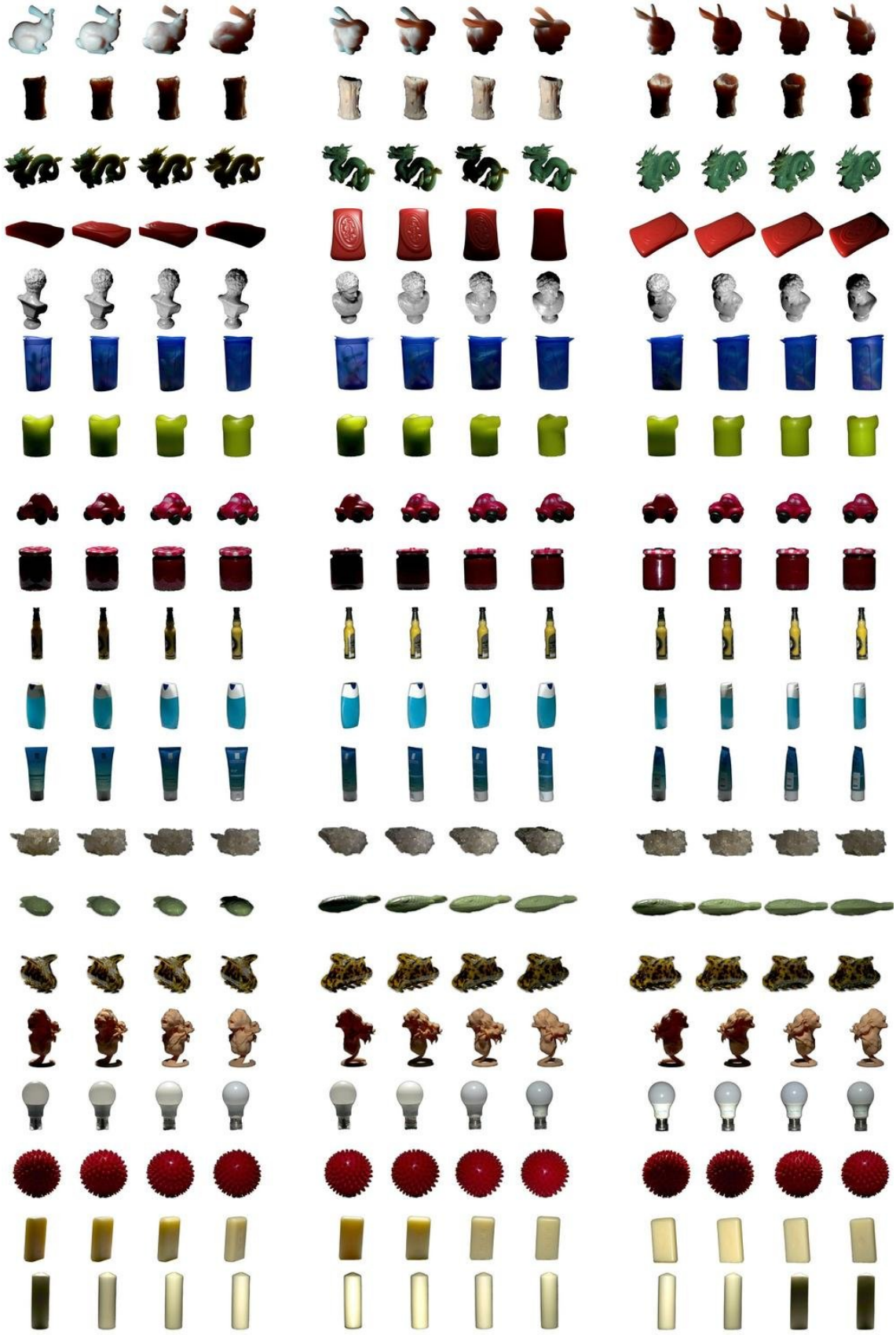


Figure 8. Dataset overview showing synthetic (top) and real (bottom) objects under different illumination and viewpoints.

Figure 9. Overview of OLAT capture setup. A hemispherical lighting rig sequentially illuminates the object with one light at a time, capturing per-light images from fixed viewpoints.

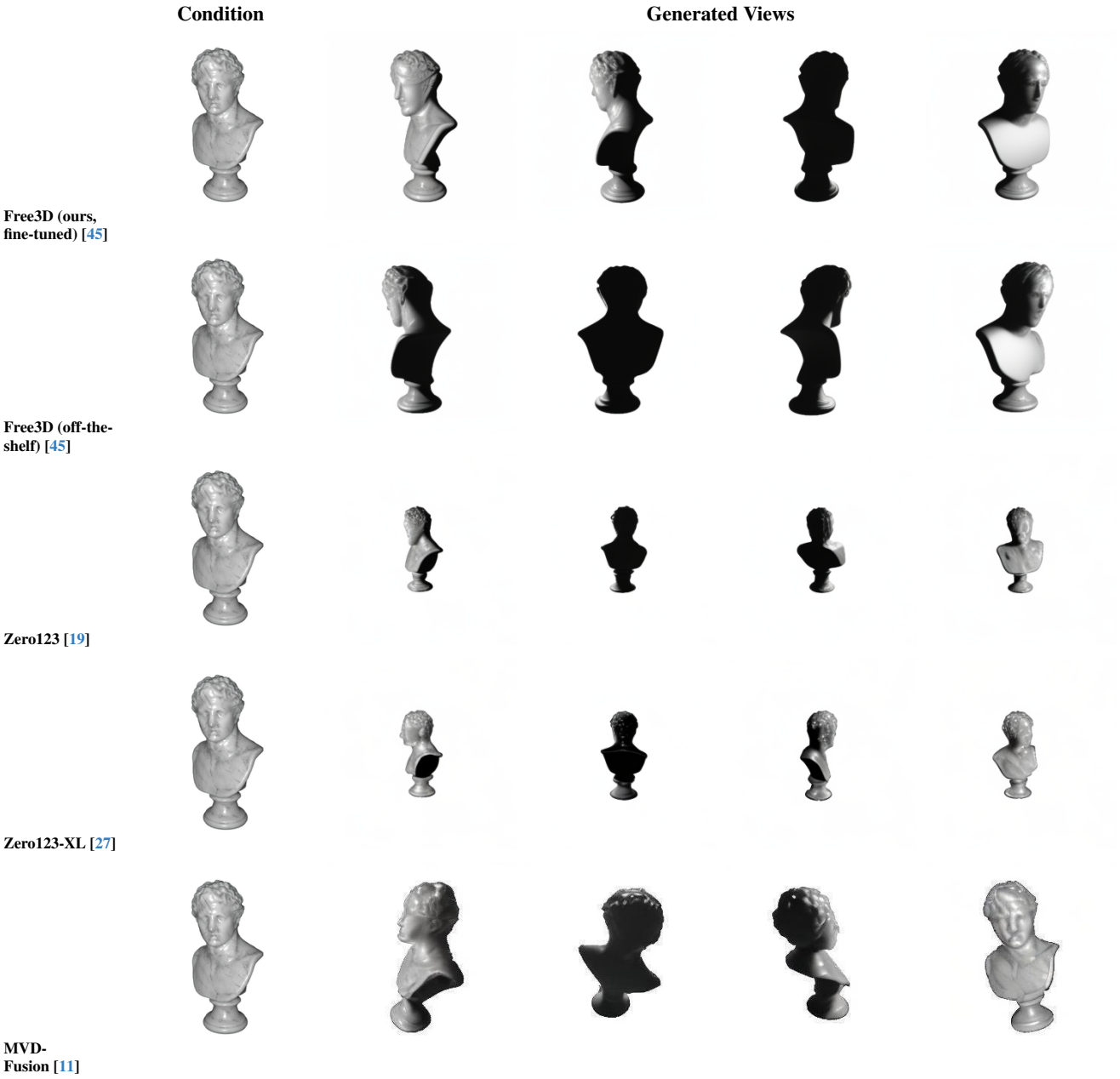


Figure 10. **Multi-view diffusion: comparison on a synthetic "statue" object.** Each row shows novel views generated from a single conditioning image (first column) by different NVS models. Our Free3D fine-tuning better preserves object identity and silhouette coherence than off-the-shelf Free3D and improves geometric consistency over other baselines.

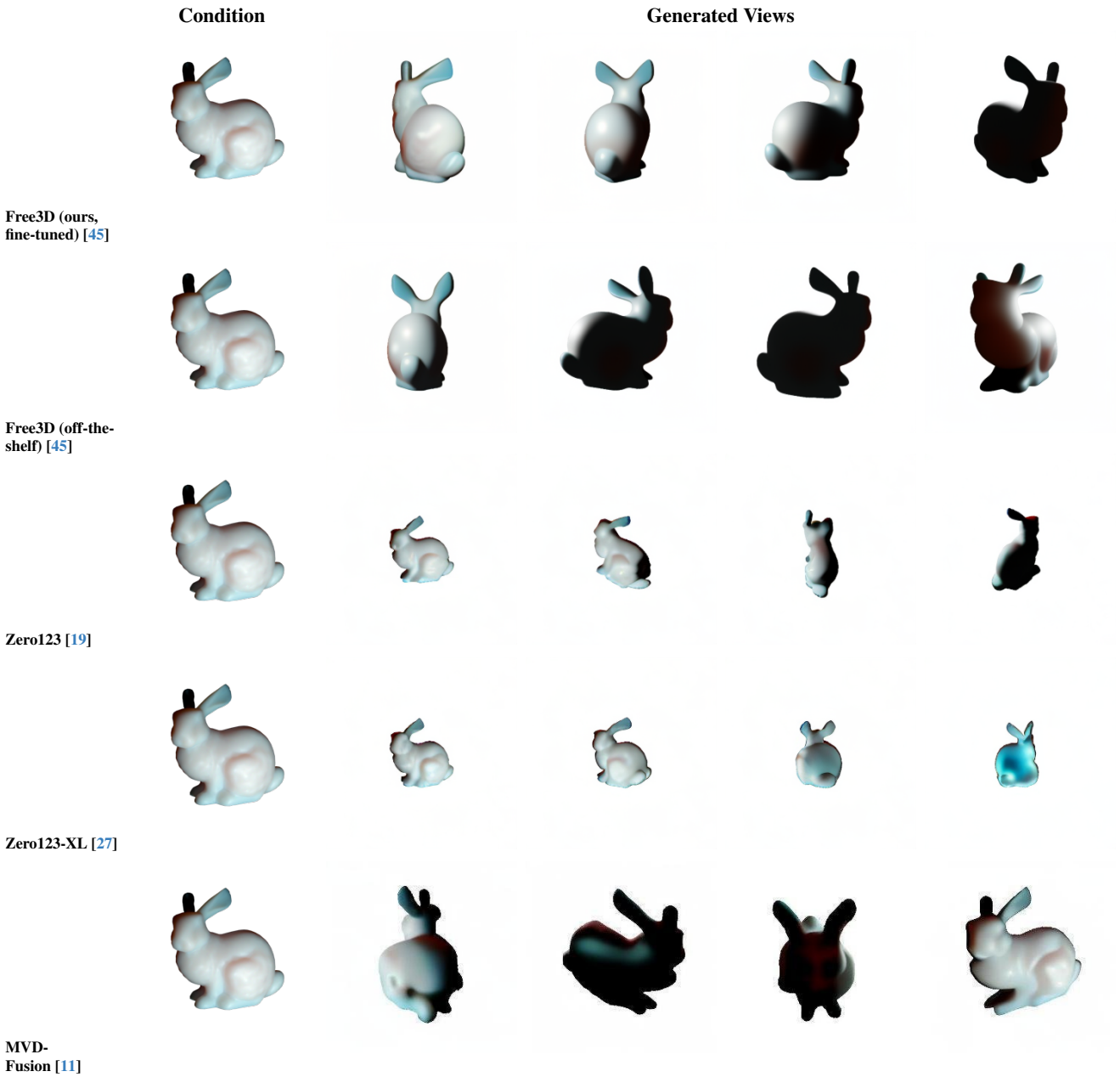


Figure 11. **Multi-view diffusion on a real "bunny" object.** Comparison between our fine-tuned Free3D model and several off-the-shelf NVS baselines. Fine-tuning improves viewpoint coherence and reduces geometric distortions while preserving appearance across views.

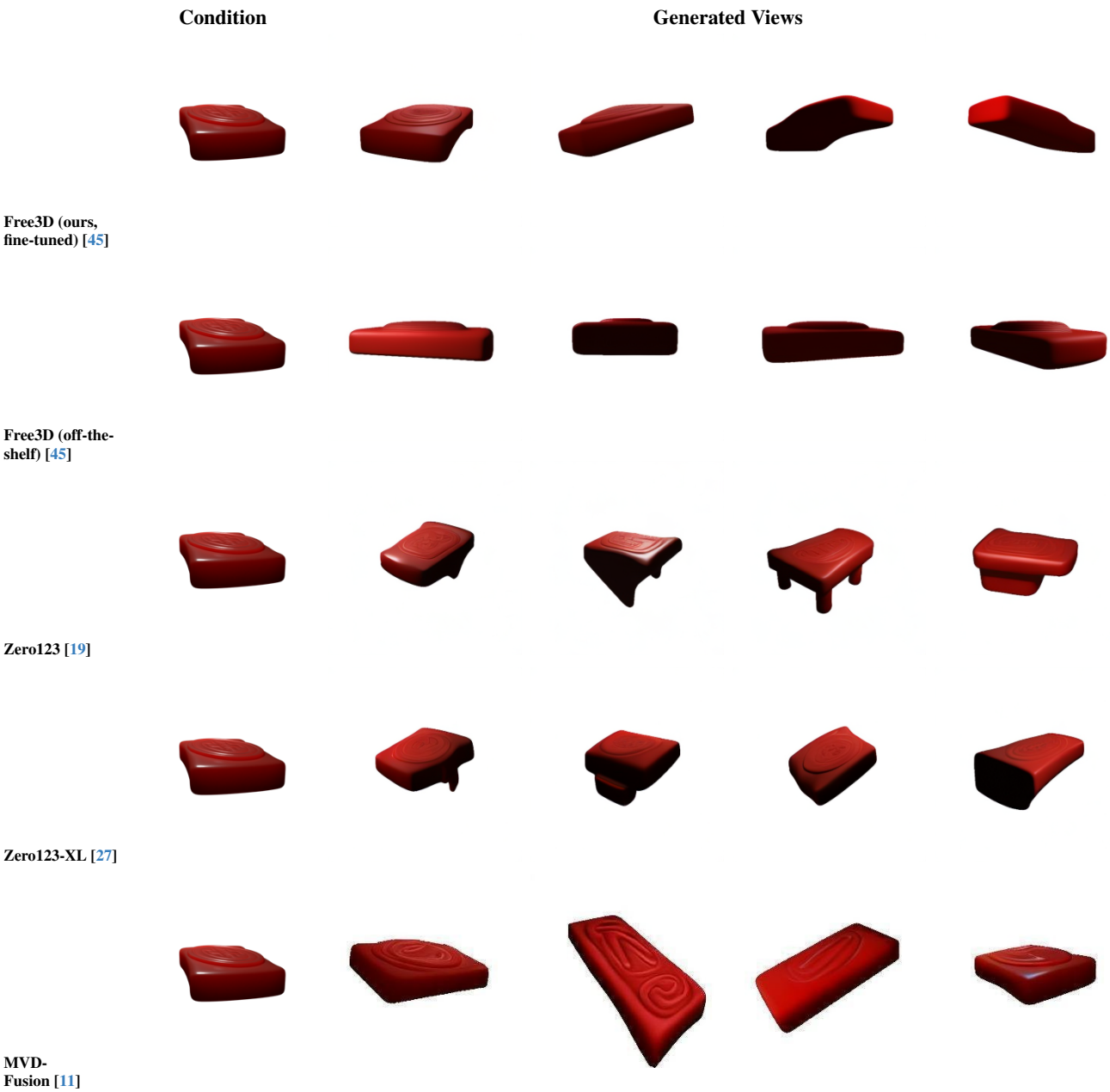


Figure 12. **Multi-view diffusion on a translucent "soap" object.** Our fine-tuned Free3D model better respects the global shape and translucency cues across viewpoints than off-the-shelf Free3D and other NVS baselines, while remaining suitable as an augmentation prior for DIAMOND-SSS.

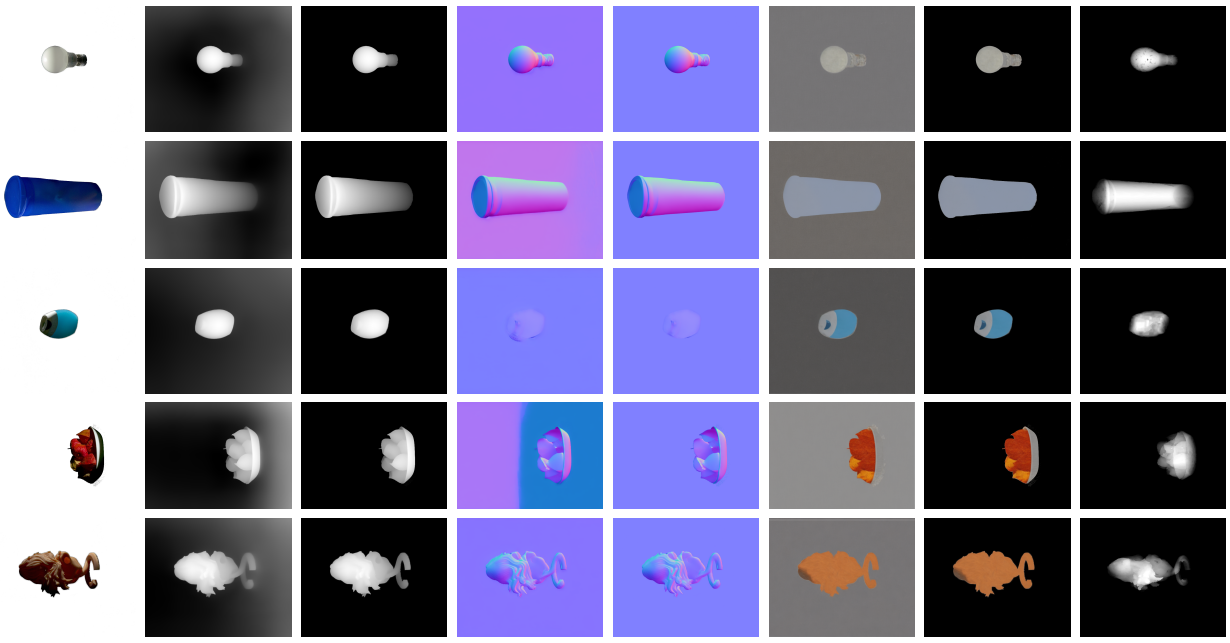


Figure 13. **Samples illustrating the compositing conventions for different conditions.** Each row shows the input RGB image and its corresponding map before and after applying the foreground mask.

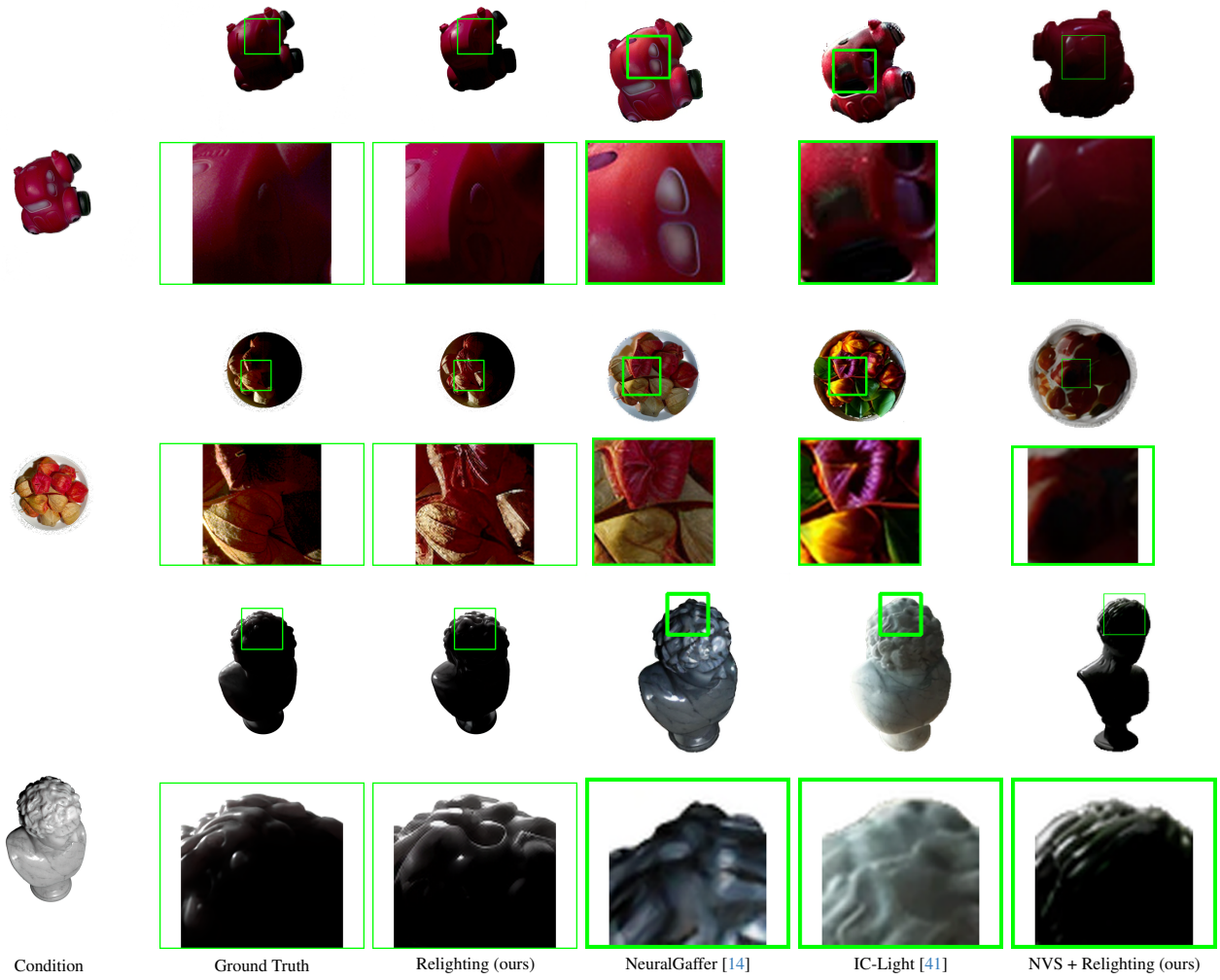


Figure 14. **Qualitative relighting comparison.** For three representative objects (two real, one synthetic), we compare our relighting-only model and our full NVS+relighting pipeline against NeuralGaffer [14] and IC-Light [41] under matched target illumination. Our method better preserves geometry and material appearance, and produces more consistent subsurface scattering cues (e.g., soft shadows and glow) across diverse lighting conditions.

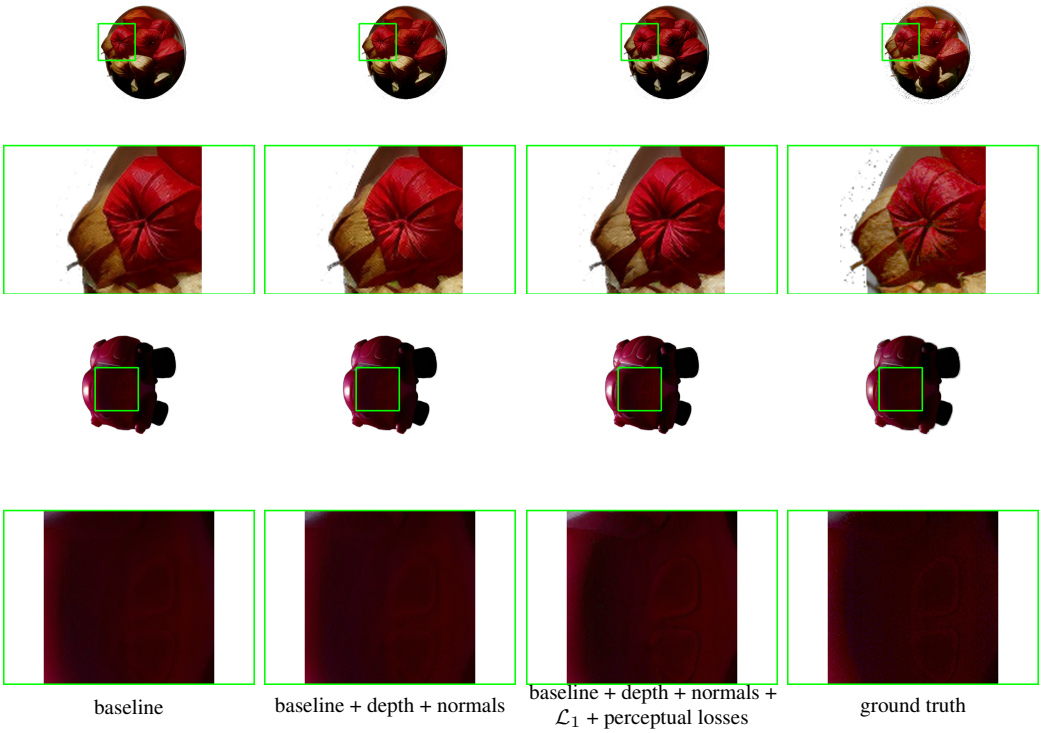


Figure 15. **Relighting ablation.** Effect of progressively adding geometric conditioning (depth, normals) and pixel/perceptual losses to the relighting diffusion model. Additional conditioning and perceptual supervision improve shadow placement, reduce halo artifacts, and better reproduce the soft scattering patterns of translucent materials.

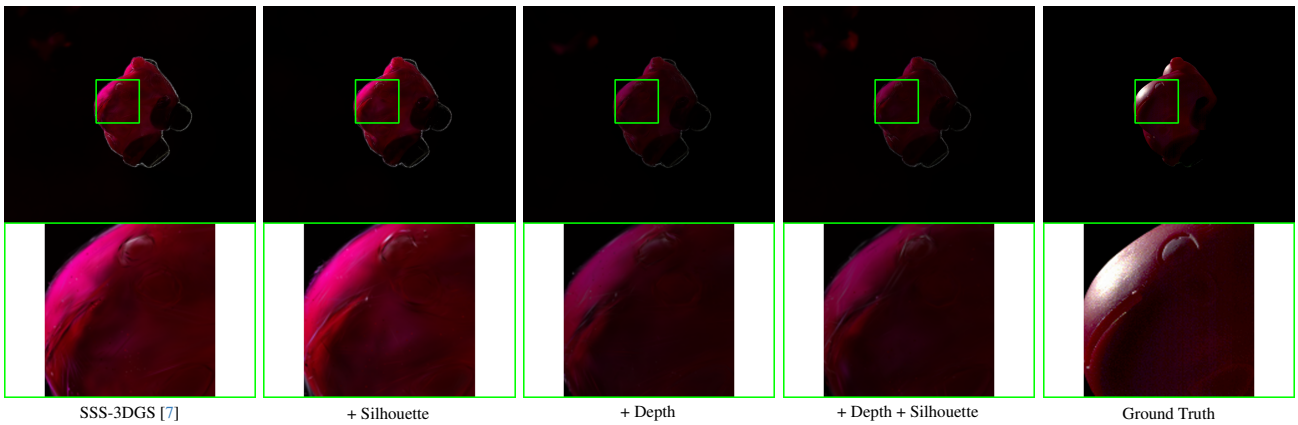
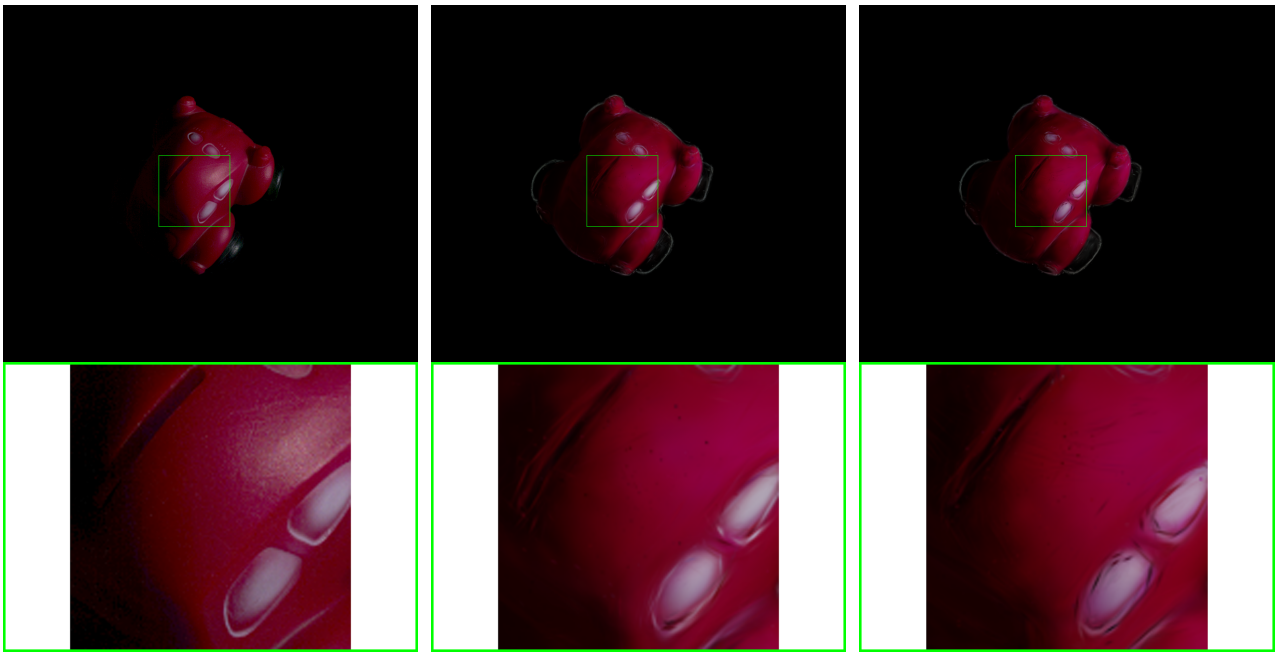
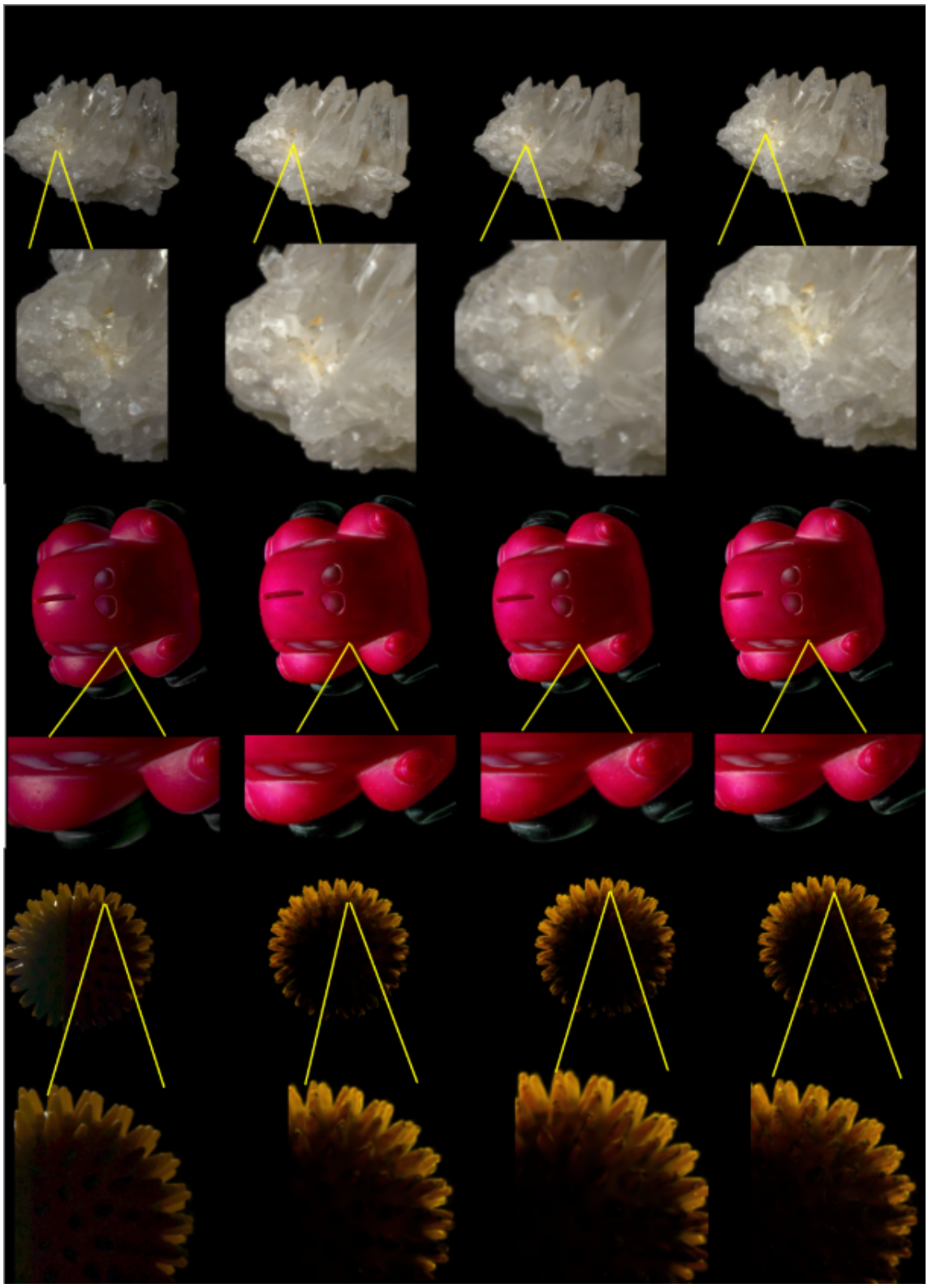


Figure 16. **Synthetic-data ablation of multi-view geometric losses.** Silhouette and depth consistency each improve boundary stability and metric alignment, with the combination yielding the most coherent results.



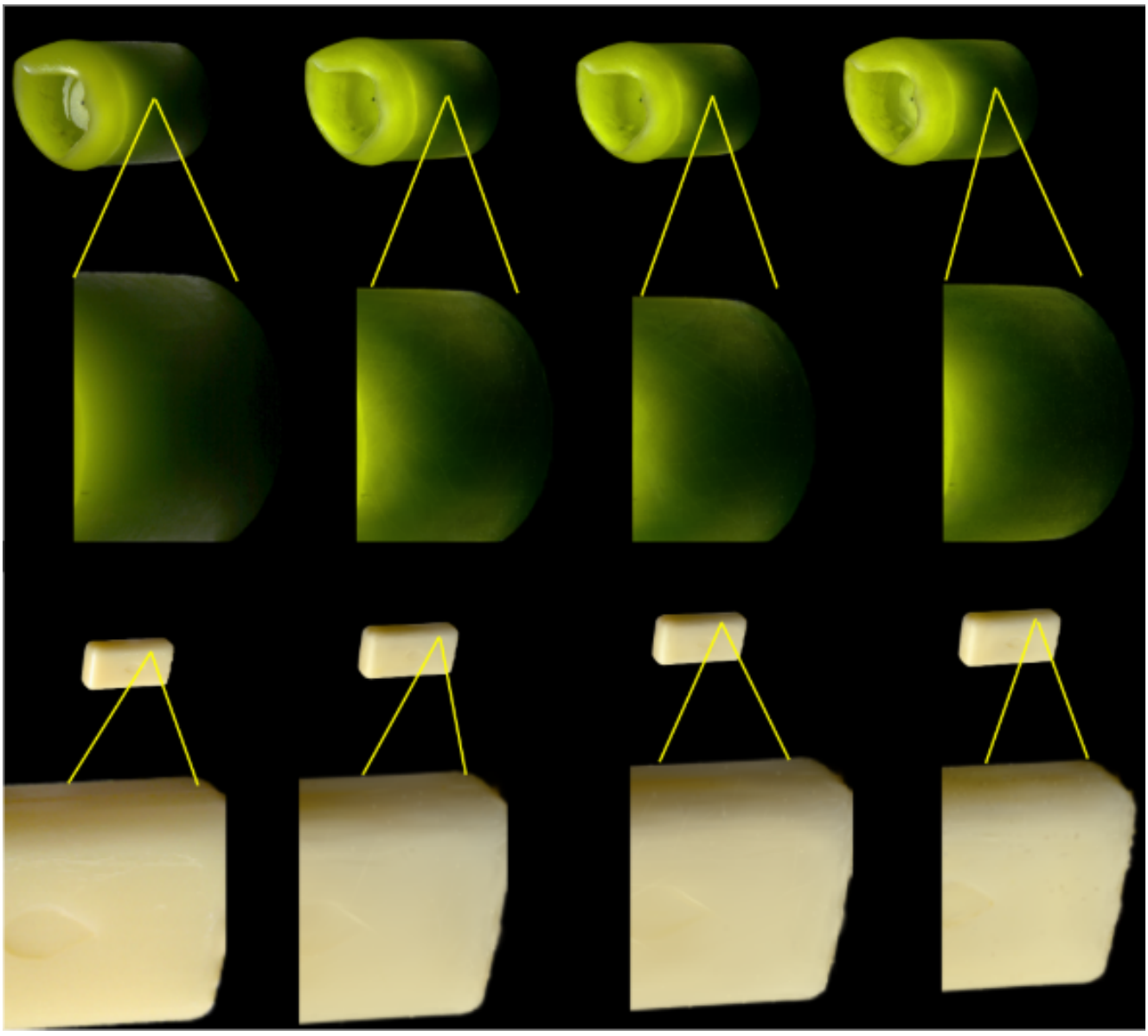
Ground Truth | Ours | SSS-3DGS [7] + Synthetic augmentation

Figure 17. **Real-data reconstruction comparison.** Multi-view geometric losses significantly improve boundary correctness and reduce view-dependent distortions. In this images we are showing the full sampling setting.



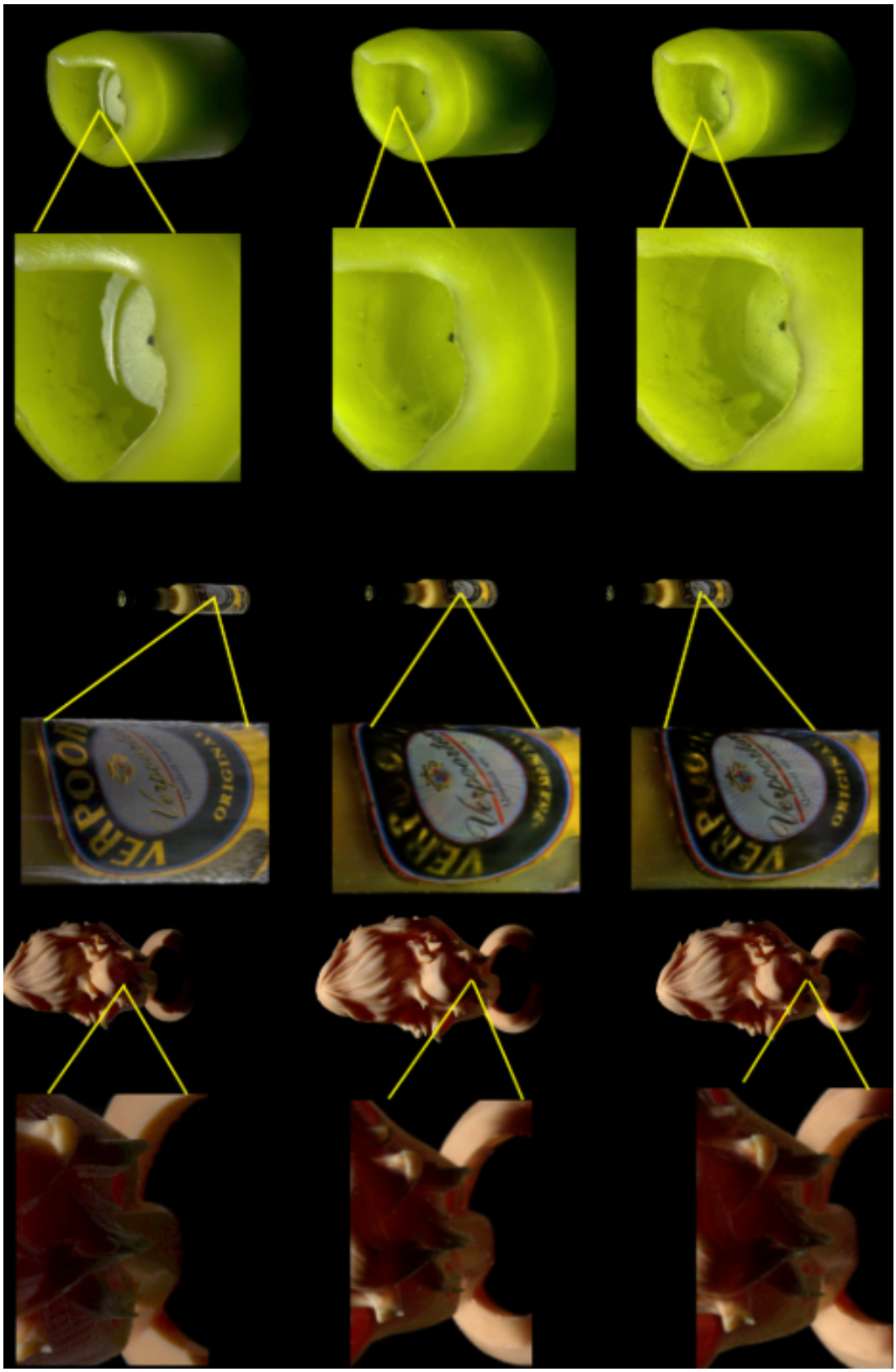
(a) Ground Truth **(b)** SSS [7] Tuned **(c)** Ours **(d)** SSS 3DGS [7] + Losses

Figure 18. **Multi-view geometric losses on real objects.** The full multi-view supervision produces sharper boundaries and reduces positional drift. Here we also compare the original method with our tuned hyperparemeters.



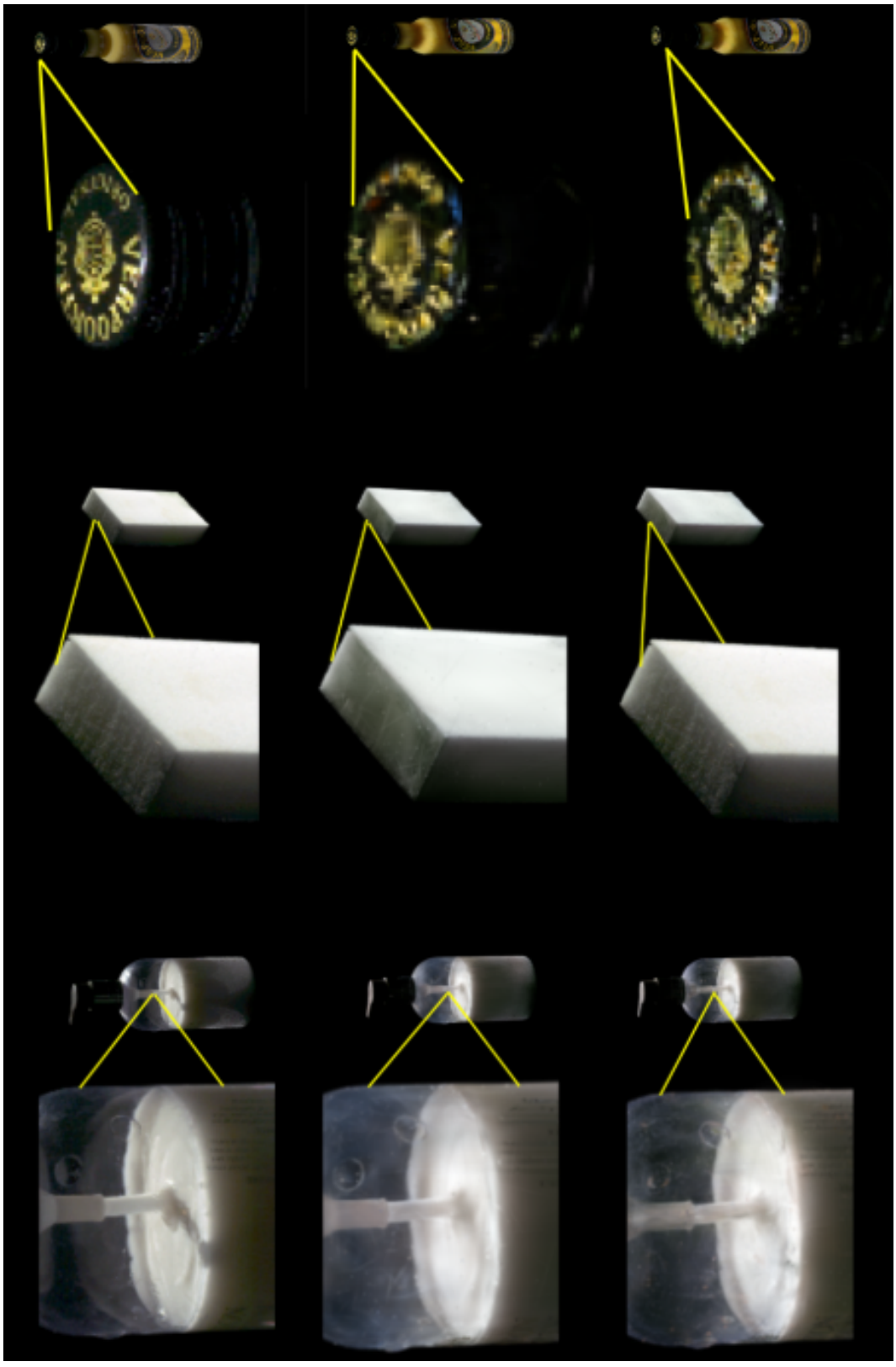
(a) Ground Truth **(b)** SSS 3DGS [7] Tuned **(c)** Ours **(d)** SSS 3DGS [7] Multi-View

Figure 19. **Synthetic vs. tuned multi-view results.** Combining parameter tuning with multi-view losses yields the most stable reconstructions. Here we also compare the original method with our tuned hyperparemeters.



(a) Ground Truth **(b)** Ours **(c)** SSS 3DGS [7] Tuned

Figure 20. **Silhouette loss effect on synthetic data.** The addition of silhouette regularization reduces boundary bleeding and sharpens object contours.



(a) Ground Truth **(b)** SSS 3DGS [7] **(c)** Ours

Figure 21. **Silhouette loss effect on synthetic data.** The addition of silhouette regularization reduces boundary bleeding and sharpens object contours.