Initial Optimization The optimization of $\mathcal{G}_c$ incorporates color, mask, and monocular depth losses. The color loss combines L1 and D-SSIM losses from 3D Gaussian Splatting:

$$\mathcal{L}_1 = \|x - x^{\text{ref}}\|_1, \quad \mathcal{L}_{\text{D-SSIM}} = 1 - \text{SSIM}(x, x^{\text{ref}}), \tag{2}$$

where x is the rendering and x^{ref} is the corresponding reference image. A binary cross entropy (BCE) loss [Jadon 2020] is applied as mask loss:

$$\mathcal{L}_{\text{m}} = -(m^{\text{ref}} \log m + (1 - m^{\text{ref}}) \log(1 - m)), \tag{3}$$

where m denotes the object mask. A shift and scale invariant depth loss is utilized to guide geometry:

$$\mathcal{L}_{\text{d}} = \|D^* - D^*_{\text{pred}}\|_1, \tag{4}$$

where D^* and D^*_{pred} are per-frame rendered depths and monocularly estimated depths [Bhat et al. 2023] respectively. The depth values are computed following a normalization strategy [Ranftl et al. 2020]:

$$D^* = \frac{D - \text{median}(D)}{\frac{1}{M}\sum_{i=1}^{M} |D - \text{median}(D)|}, \tag{5}$$

where M denotes the number of valid pixels. The overall loss combines these components:

$$\mathcal{L}_{\text{ref}} = (1 - \lambda_{\text{SSIM}})\mathcal{L}_1 + \lambda_{\text{SSIM}}\mathcal{L}_{\text{D}-\text{SSIM}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} + \lambda_{\text{d}}\mathcal{L}_{\text{d}}, \tag{6}$$

where λ_{SSIM}, λ_{m}, and λ_{d} control the magnitude of each term. Thanks to the efficient initialization, our training speed is remarkably fast. It only takes 1 minute to train a coarse Gaussian representation $\mathcal{G}_c$ at a resolution of 779×520.