PoInit-of-View: Poisoning Initialization of Views Transfers Across Multiple 3D Reconstruction Systems

Supplementary Material

This appendix provides additional theoretical analysis and derivation, more empirical proof, and ablation results for PoInit-of-View. It is organized into two main sections:

- Section A provides the proofs of theoretical analysis.
- Section B provides more ablations of hyperparameters and defense discussion.

A. Proofs of Theoretical Results

This section establishes how pixel-level perturbations propagate through cross-view gradients, descriptor representations, and ultimately the robustness of Structure-from-Motion (SfM).

A.1. From Cross-View Inconsistency to Descriptor Divergence

Let R_i be a clean reference image and A_j a poisoned one with perturbation δ. Applying the Sobel operator gives

$$G(A_j) = G(R_j + \delta) = G(R_j) + J\delta, \tag{12}$$

where J is the Sobel Jacobian. Under bounded-smoothness,

$$\|G(R_i) - G(R_j)\|_1 \le \tau_g, \tag{13}$$

and since Sobel is linear,

$$c_s^{\min}\|\delta\|_1 \le \|J\delta\|_1 \le c_s^{\max}\|\delta\|_1. \tag{14}$$

Thus,

$$\tau_g + c_s^{\min}\|\delta\|_1 \;\le\; \mathrm{LCVI} \;\le\; \tau_g + c_s^{\max}\|\delta\|_1. \tag{15}$$

Local descriptors such as SIFT/RootSIFT are piecewise linear under fixed orientation-bin assignments. Therefore, within a small radius $r > 0$, the descriptor mapping is bi-Lipschitz:

$$L_r^{\min}\|G_1 - G_2\|_1 \le \|\phi(G_1) - \phi(G_2)\|_2 \le L_r^{\max}\|G_1 - G_2\|_1. \tag{16}$$

Since $\|\delta\|_\infty \le \epsilon$, the perturbed gradients remain in this local region, giving

$$\|\phi(G(R_i)) - \phi(G(A_j))\|_2 \;\ge\; L_r^{\min}\,\mathrm{LCVI} = \beta_r\,\mathrm{LCVI}. \tag{17}$$

A.2. From Descriptor Divergence to Matching Probability

Empirically (see Figure 9), the descriptor deviation $D = \|\phi(G_1) - \phi(G_2)\|_2$ follows a sub-exponential tail, so for some $\alpha > 0$,

$$\Pr[D < \tau_d] \le \exp(-\alpha D). \tag{18}$$

Combining this with the previous lower bound $D \ge \beta_r\Delta$ (where $\Delta = \mathrm{LCVI} - \tau_g$) yields

$$p_{\mathrm{match}} \le \exp(-\alpha\beta_r\Delta), \tag{19}$$

showing that correspondence reliability decays exponentially with cross-view inconsistency.

Let the inlier count be $M \sim \mathrm{Binomial}(N, p_{\mathrm{match}})$. Chernoff bound states that for any $0 < \epsilon < 1$,

$$\Pr[M \le (1-\epsilon)Np_{\mathrm{match}}] \le \exp\!\left(-\tfrac{1}{2}\epsilon^2 Np_{\mathrm{match}}\right). \tag{20}$$

Hence the failure probability satisfies

$$\Pr[\eta < \eta_{\min}] \ge 1 - \exp\!\left(-\tfrac{1}{2}\epsilon^2 Np_{\mathrm{match}}\right). \tag{21}$$

A.3. From Matching Probability to SfM Breakdown

SfM accepts a view-pair only if $\eta \ge \eta_{\min}$. If a critical pose-graph edge satisfies

$$\|G(R_i) - G(A_j)\|_1 \ge \tau_g + \frac{\tau_d}{\beta_r} + \Delta, \tag{22}$$

then its matching probability obeys the exponential decay in Section B.2.

Using the union bound over m critical edges (independence not required), the probability that at least one edge fails is

$$\Pr[\text{SfM fails}] \ge 1 - m\,\exp\!\left(-\tfrac{1}{2}\epsilon^2 Np_{\mathrm{match}}\right). \tag{23}$$

Thus, SfM fails with overwhelming probability whenever the cross-view inconsistency exceeds

$$L_{\mathrm{th}} = \tau_g + \frac{\tau_d}{\beta_r}. \tag{24}$$

This threshold closely matches the empirical collapse point observed in Figure 9 of the main paper. We provide the empirical estimation of $(\beta_r, \tau_d, \tau_g)$ and the resulting threshold L_{th} in Section B.2.

B. Additional Experiments

B.1. Complete Results on All Datasets

Table 6 and Table 7 illustrate the distinct impact of PoInit-of-View attacks on synthetic versus real-world scenes. In NeRF-Synthetic Table 6, scenes are rendered under fully controlled conditions with consistent illumination, texture, and geometry. Feature correspondences are stable and the overall parallax is limited, making SfM relatively robust.

Table 6. Clean vs. Poisoned comparison across reconstruction pipelines on *Blender (NeRF-Synthetic)* with $\rho = 16/255$. Clean results in black; poisoned results in parentheses.

Scene	Colmap			Instant NGP			Mip-Splatting		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
lego	12.31 (10.52)	0.776 (0.688)	0.220 (0.276)	35.65 (31.23)	0.981 (0.894)	0.020 (0.024)	35.45 (32.10)	0.982 (0.910)	0.021 (0.025)
drums	8.06 (6.90)	0.657 (0.579)	0.296 (0.364)	24.57 (21.46)	0.930 (0.842)	0.109 (0.131)	26.14 (23.33)	0.953 (0.882)	0.046 (0.055)
ficus	15.12 (12.86)	0.838 (0.744)	0.143 (0.178)	30.29 (26.58)	0.972 (0.895)	0.031 (0.037)	35.12 (31.53)	0.988 (0.923)	0.013 (0.016)
hotdog	12.21 (10.37)	0.831 (0.744)	0.195 (0.234)	37.02 (32.63)	0.982 (0.900)	0.037 (0.045)	37.78 (33.72)	0.985 (0.915)	0.028 (0.033)
materials	14.56 (12.36)	0.802 (0.708)	0.193 (0.238)	28.96 (25.15)	0.944 (0.861)	0.069 (0.083)	30.12 (26.64)	0.960 (0.893)	0.044 (0.052)
mic	9.29 (7.99)	0.765 (0.686)	0.182 (0.221)	35.41 (31.23)	0.989 (0.900)	0.016 (0.019)	35.55 (31.84)	0.991 (0.917)	0.008 (0.010)
ship	10.02 (8.62)	0.616 (0.548)	0.335 (0.420)	30.61 (26.66)	0.892 (0.816)	0.136 (0.164)	30.78 (27.39)	0.904 (0.843)	0.132 (0.158)
chair	15.42 (13.01)	0.847 (0.764)	0.148 (0.179)	35.07 (30.58)	0.984 (0.900)	0.023 (0.028)	35.70 (31.71)	0.986 (0.914)	0.018 (0.022)
Average	**12.12** (10.06)	**0.766** (0.684)	**0.214** (0.274)	**32.20** (28.19)	**0.959** (0.891)	**0.055** (0.067)	**33.33** (29.46)	**0.969** (0.903)	**0.039** (0.046)
Avg. Drop (%)	-17.0 %	-10.7 %	+28.0 %	-12.4 %	-7.1 %	+21.4 %	-11.6 %	-6.8 %	+19.0 %

Table 7. Clean vs. Poisoned comparison across reconstruction pipelines on *Mip-NeRF 360 Dataset* with $\rho = 16/255$. Clean results in black; poisoned results in parentheses.

Scene	Colmap			Instant NGP			Mip-Splatting		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
garden	18.87 (14.66)	0.468 (0.411)	0.477 (0.566)	24.78 (19.70)	0.654 (0.587)	0.346 (0.435)	27.47 (22.52)	0.869 (0.794)	0.124 (0.199)
bicycle	18.29 (14.21)	0.352 (0.309)	0.644 (0.764)	23.00 (18.30)	0.526 (0.472)	0.489 (0.616)	25.25 (20.71)	0.765 (0.698)	0.243 (0.389)
flowers	14.50 (11.27)	0.257 (0.226)	0.634 (0.753)	20.54 (16.32)	0.462 (0.414)	0.478 (0.602)	21.60 (17.71)	0.605 (0.553)	0.371 (0.594)
treehill	15.73 (12.22)	0.340 (0.299)	0.726 (0.862)	22.36 (17.78)	0.528 (0.473)	0.526 (0.662)	22.65 (18.57)	0.633 (0.578)	0.381 (0.611)
stump	19.65 (15.27)	0.366 (0.322)	0.646 (0.767)	23.84 (18.95)	0.590 (0.530)	0.439 (0.552)	26.64 (21.85)	0.774 (0.707)	0.251 (0.402)
kitchen	17.35 (13.48)	0.497 (0.437)	0.575 (0.683)	29.02 (23.07)	0.844 (0.757)	0.255 (0.321)	31.25 (25.63)	0.926 (0.845)	0.155 (0.248)
bonsai	14.06 (10.92)	0.548 (0.482)	0.586 (0.696)	30.30 (24.09)	0.890 (0.798)	0.295 (0.371)	31.96 (26.21)	0.941 (0.860)	0.254 (0.407)
counter	15.04 (11.69)	0.549 (0.483)	0.520 (0.617)	26.56 (21.13)	0.812 (0.728)	0.373 (0.469)	29.04 (23.81)	0.907 (0.828)	0.258 (0.413)
room	16.55 (12.86)	0.628 (0.552)	0.505 (0.599)	29.16 (23.19)	0.850 (0.764)	0.383 (0.482)	31.54 (25.86)	0.918 (0.839)	0.286 (0.458)
Average	**16.67** (13.11)	**0.445** (0.391)	**0.590** (0.700)	**25.51** (20.27)	**0.684** (0.612)	**0.398** (0.501)	**27.49** (22.54)	**0.815** (0.744)	**0.258** (0.414)
Avg. Drop (%)	-22.3 %	-12.1 %	+18.7 %	-20.5 %	-10.3 %	+25.9 %	-18.0 %	-8.7 %	+60.2 %

Consequently, although PoInit-of-View introduces cross-view inconsistencies, all reconstruction pipelines exhibit only moderate degradation in PSNR/SSIM and a small increase in LPIPS.

In contrast, Mip-NeRF360 Table 7 contains real photographic data with large-baseline viewpoints, complex textures, and significant illumination variations. These factors naturally weaken feature stability and make SfM more susceptible to matching errors. Under such conditions, the cross-view perturbations injected by PoInit-of-View are amplified during triangulation and bundle adjustment, resulting in larger pose drift and geometric distortion. This leads to substantially more severe performance drops across all pipelines, as reflected by stronger declines in PSNR/SSIM and pronounced increases in LPIPS. Overall, the two tables jointly demonstrate that while PoInit-of-View has limited influence in idealized synthetic settings, its impact is significantly amplified in real, visually complex environments, causing much stronger degradation in both SfM and neural rendering pipelines. As shown in Figure 10, to estimate the upper bound of our method on SfM statistics, we manually mask keypoints and remove all pixel-level structural features.

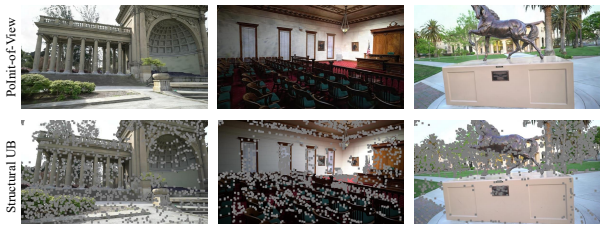

Figure 10. Comparison between regular PoInit-of-View inputs (top row) and our Structural Upper Bound setting (bottom row). In the Structural Upper Bound setting, we manually mask keypoints and remove all pixel-level structural features to estimate the upper bound of our method on SfM statistics. This produces purely structure-free imagery, isolating the impact of structural cues on reconstruction.

B.2. Estimation of the Breakdown Threshold

We estimate the breakdown threshold L_{th}

$$L_{\text{th}} = \tau_g + \frac{\tau_d}{\beta_r}. \tag{25}$$

purely from statistics of the clean SfM reconstruction.

Step 1: Estimating τ_g. From the clean COLMAP reconstruction, we collect all verified keypoint correspondences (p_i, p_j) on the Tanks & Temples benchmark. For each correspondence, we extract aligned patches around p_i and p_j and compute Sobel gradients $G(I_i(p_i))$ and $G(I_j(p_j))$. We measure the gradient discrepancy

$$d_g = \|G(I_i(p_i)) - G(I_j(p_j))\|_1 , \tag{26}$$

and define τ_g as the empirical mean of d_g over all clean correspondences.

Step 2: Estimating τ_d. Using the same set of correspondences, we compute descriptor distances

$$d_d = \|\phi(I_i(p_i)) - \phi(I_j(p_j))\|_2 . \tag{27}$$

We sort all d_d and set τ_d to the 5-th percentile of this distribution, which yields a conservative descriptor tolerance under clean matching.

Step 3: Estimating β_r. We randomly sample K clean patch pairs (G_1, G_2) and compute

$$r = \frac{\|\phi(G_1) - \phi(G_2)\|_2}{\|G_1 - G_2\|_1}. \tag{28}$$

We then define

$$\beta_r = \mathbb{E}[r]. \tag{29}$$

Finally, we plug the three quantities into

$$L_{\text{th}} = \tau_g + \frac{\tau_d}{\beta_r}. \tag{30}$$

The predicted SfM breakdown threshold

$$L_{\text{th}} = \tau_g + \tau_d/\beta_r$$

evaluates to ≈ 0.26 on Tanks & Temples as shown in Figure 9. We provide the pseudo-code for the estimation of the breakdown threshold in Algorithm 2.

Algorithm 2 Estimating τ_g, τ_d, β_r, and L_{th} from clean data

Require: Clean input images $\{I_i\}$; COLMAP verified correspondences $\mathcal{M}$
Ensure: Estimated τ_g, τ_d, β_r, and final threshold L_{th}
Step 1: Initialize statistics containers
$\mathcal{G} \leftarrow [\,]$ {Stores all gradient differences d_g}
$\mathcal{D} \leftarrow [\,]$ {Stores all descriptor differences d_d}
$\mathcal{R} \leftarrow [\,]$ {Stores all ratios $r = d_d/d_g$}
Step 2: Process each verified correspondence
for all $(i, j, k) \in \mathcal{M}$ **do**
  Extract aligned patches $P_i^{(k)}$, $P_j^{(k)}$ (size 11×11) {Local neighborhoods around matched keypoints}
  Compute Sobel gradients: $G_i^{(k)} = G(P_i^{(k)})$, $G_j^{(k)} = G(P_j^{(k)})$
  $d_g^{(k)} = \|G_i^{(k)} - G_j^{(k)}\|_1$
  Append $d_g^{(k)}$ to $\mathcal{G}$
  $d_d^{(k)} = \|\phi_i^{(k)} - \phi_j^{(k)}\|_2$
  Append $d_d^{(k)}$ to $\mathcal{D}$
  if $d_g^{(k)} > 10^{-4}$ **then**
    Compute ratio $r^{(k)} = d_d^{(k)}/d_g^{(k)}$
    Append $r^{(k)}$ to $\mathcal{R}$
  end if
end for
Step 3: Aggregate statistics
Remove top 1% of $\mathcal{G}$ {Filter out rare extreme gradients}
$\tau_g \leftarrow \text{mean}(\mathcal{G})$ {Cross-view gradient level}
Sort $\mathcal{D}$
$\tau_d \leftarrow \text{Quantile}_{0.05}(\mathcal{D})$ {Descriptor tolerance}
Remove top 5% of $\mathcal{R}$
$\beta_r \leftarrow \text{mean}(\mathcal{R})$ {Descriptor–gradient sensitivity}
Step 4: Compute theoretical SfM breakdown threshold
$L_{\text{th}} \leftarrow \tau_g + \tau_d/\beta_r$
return $\tau_g,\ \tau_d,\ \beta_r,\ L_{\text{th}}$

B.3. More Ablations for Hyperparameters

Ablations for T. As shown in Figure 11-(a), increasing the PGD steps T strengthens the attack only up to $T \approx 15$. Beyond this point, both $\mathcal{L}_{\text{CVI}}$ and the collapse ratio saturate, indicating that PoInit-of-View does not rely on excessive optimization steps but rather on crossing the theoretical threshold L_{th}.

Ablations for α. As shown in Figure 11-(b), we sweep the PGD step size $\alpha \in \{0.5, 1, 2, 4\}/255$. Small step sizes under-optimize the inconsistency objective and lead to limited changes in cross-view gradients. Moderate values (e.g., $\alpha = 2/255$) maximize the growth of $\mathcal{L}_{\text{CVI}}$ and yield the strongest collapse. Larger step sizes cause optimization overshoot and introduce high-frequency artifacts without improving cross-view inconsistency.

Ablations for r. As shown in Figure 11-(c), we vary the poisoning ratio $r \in \{0.1, 0.3, 0.6, 1.0\}$. When only 10% of views are poisoned, SfM remains largely stable. At $r = 0.3$, we observe noticeable degradation, while $r = 0.6$ (our default) offers the best trade-off between attack effectiveness and perceptual quality. Increasing to $r = 1.0$ yields only marginal additional gain, demonstrating that PoInit-of-View does not require poisoning all input views.

B.4. Additional Low-Cost Defenses

Structure Randomization (Patch Masking). Randomly replace ∼5% of 16×16 patches with local blur; helps but may reduce keypoint density.

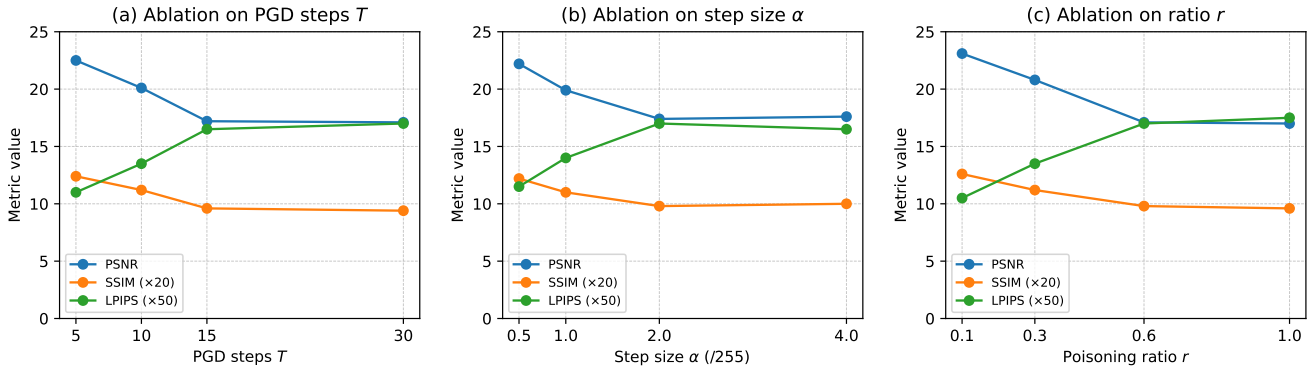

Figure 11. Ablation studies on hyperparameters T, α, and r. Higher PSNR/SSIM indicates better reconstruction; lower LPIPS indicates more perceptual degradation.

Table 8. Extended defense comparison on Tanks & Temples (T&T). Higher values indicate better defense. ***PoInit-of-View*** remains effective against defenses.

Defense	**Reg. Imgs** ↑	**Triang. Pts (k)** ↑	**3D Pts (M)** ↑
None	11.2	4.8	0.10
Patch Masking	28.5	13.2	0.23
JPEG (q=85)	36.9	17.0	0.29
Color Jitter	22.7	10.1	0.18
Clean	87.3	67.6	2.37

Photometric Normalization (Color Jitter). Small random brightness/contrast/saturation/hue; modest gains mostly against photometric-only poisons.

Table 8 shows that ***PoInit-of-View*** significantly degrades SfM, reducing registered views to only 11.2 without any protection. Among lightweight defenses, JPEG compression is the most effective, nearly tripling the number of registered images by suppressing the high-frequency cross-view inconsistencies introduced by our attack. Patch Masking also improves robustness by weakening corrupted local structures, while Color Jitter provides only modest gains since PoInit-of-View primarily exploits geometric, rather than photometric, cues. Overall, simple preprocessing can partially mitigate PoInit-of-View, but a substantial gap remains compared to clean performance.

B.5. Modern Learned Descriptors

As shown in Table 9, to verify the cross-model generalizability of our attack, we extend our evaluation beyond the standard SIFT (Colmap) initializer to modern learned descriptors. We focus on SIFT (Colmap) as it is widely used as a standard SfM initializer in many reconstruction pipelines. Additionally, we replace SIFT with two well-known learned descriptors, SuperPoint+SuperGlue (SP+SG) [34] and LoFTR [39], via the widely used hierarchical localization toolbox (HLOC), while keeping the remaining SfM backend unchanged. Results on all T&T below confirm that our attack still yields substantial degradation, although learned descriptors are more robust than SIFT.

Table 9. Effect of poisoning on different descriptors in SfM reconstruction.

Descriptors	Setting	Reg.%	Triang.(k)	3D Pts(M)	Collapse
SIFT	Clean/Poi.	93.5/24.3	73.6/13.7	3.07/0.41	0.88
SP+SG	Clean/Poi.	95.8/46.7	81.4/31.2	3.32/1.05	0.62
LoFTR	Clean/Poi.	96.6/54.9	84.9/36.8	3.45/1.22	0.55

B.6. Collapse Ratio

As shown in Figure 12, to provide an intuitive understanding of our proposed evaluation metric, we conduct a qualitative ablation on the Collapse Ratio. Figure 12 visualizes the 3D reconstruction outcomes across varying collapse ratios (0.15, 0.33, and 0.81) alongside the clean baseline. As shown, there is a clear correlation between the metric and the visual integrity of the scene: a larger collapse ratio consistently translates to more severe structural degradation. Notably, at a high ratio of 0.81, the reconstruction is nearly entirely corrupted. This visualization confirms that the Collapse Ratio serves as a reliable and accurate indicator of the reconstruction quality and, consequently, the effectiveness of the attack.

Figure 12. Visual effect of different collapse ratios on 3D reconstruction.

B.7. Theory Indeed Guides the Attack Design

Our theory proves that the attack is guaranteed to succeed when $\mathcal{L}_{\text{CVI}} > L_{th}$. It directly motivates our attack design of maximizing $\mathcal{L}_{\text{CVI}}$ (Eqs. (5)). In particular, **Appendix C.2** has derived $L_{th} \approx 0.26$ on T&T dataset. The figure below confirms that PGD with our setting of 15 steps ($\mathcal{L}_{\text{CVI}} = 0.28$) already satisfies the condition. Fig. 11(a) in **Appendix C.3.1** also confirms that the attack performance gets saturated at 15 steps.

B.8. Minimum poisoned views & view selection

As shown in Table 10, we repeat the random selection of poisoned views 5 times. Results below show very small variances, which makes sense because the reconstruction is not expected to be biased towards specific views.

Table 10. Reconstruction results under five random poisoned-view selections, showing low variance.

Poison	Reg.%	Triang.(k)	3D Pts(M)	Collapse
Rand. 5 times	24.3±0.3	13.7±1.2	0.41±0.07	0.88±0.04

B.9. Theory guides the attack design

As shown in Figure 13, our theory proves that the attack is guaranteed to succeed when $\mathcal{L}_{\text{CVI}} > L_{th}$. It directly motivates our attack design of maximizing $\mathcal{L}_{\text{CVI}}$ (Eqs. (5)). In particular, **Appendix C.2** has derived $L_{th} \approx 0.26$ on T&T dataset. The figure below confirms that PGD with our setting of 15 steps ($\mathcal{L}_{\text{CVI}} = 0.28$) already satisfies the condition. Fig. 11(a) in **Appendix C.3.1** also confirms that the attack performance gets saturated at 15 steps.

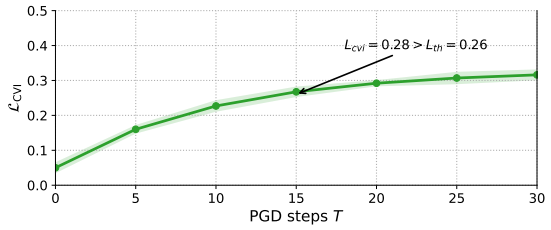

Figure 13. Cross-view inconsistency loss $\mathcal{L}_{\text{cvi}}$ vs. PGD steps T. The loss increases steadily and surpasses the threshold at $T = 15$.