TRACE: TRANSIENT RECOGNITION, AUTHENTICATION, AND COVERAGE-AWARE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in 3D Gaussian Splatting (GS) have enabled high-quality novel view synthesis from casually captured images, yet fully automatic transient object removal remains challenging. Existing methods rely on unreliable geometric cues or assume accurate transient masks are available, often removing persistent objects that belong to transient semantic categories. We present TRACE, a fully automatic pipeline that authenticates transient objects before editing. Candidate objects are detected using off-the-shelf instance segmentation and verified through multi-view geometric persistence derived from Structure-from-Motion, with cross-view semantic consistency resolving ambiguous cases. Using the authenticated masks, TRACE selects a minimal yet informative set of views for diffusion-based inpainting by jointly optimizing geometric coverage and object visibility. Experiments on challenging real-world datasets demonstrate that TRACE significantly improves transient localization and produces higher-quality Gaussian Splatting reconstructions than existing geometry- and category-based approaches.

1 INTRODUCTION

Recent advances in 3D Gaussian Splatting (GS) (Kerbl et al., 2023; Fei et al., 2024; Chen & Wang, 2024) have transformed novel view synthesis by representing scenes as explicit, learnable 3D Gaussians initialized from a Structure-from-Motion (SfM) prior (Schonberger & Frahm, 2016). Compared to implicit neural representation based novel view synthesis, i.e, Neural Radiance Fields (NeRF) (Mildenhall et al., 2021), GS offers real-time rendering, efficient optimization, and an explicit scene representation that naturally supports geometric manipulation. These properties have fueled rapid progress in applications ranging from large-scale scene reconstruction (Lin et al., 2024) and digital twins (Guo et al., 2025) to virtual and augmented reality (Zhai et al., 2025). Beyond high-quality rendering, the explicit nature of GS has also made scene editing increasingly practical, enabling operations such as object insertion, appearance modification, and object removal by propagating edits from a subset of training images to the reconstructed 3D scene (Chen et al., 2024; Wu et al., 2024b; Ye et al., 2024). Despite these advances, reliable scene editing, particularly the automatic removal of objects/transient objects from real-world captures, remains a fundamentally challenging and relatively underexplored problem (Markin et al., 2024).

Transient objects are scene elements that do not persist throughout the image acquisition process, such as pedestrians walking through a courtyard, vehicles passing through a street, or bicycles temporarily parked within the scene. Unlike the underlying static geometry, these objects are incidental to the environment and often vary across repeated captures of the same location. Their presence introduces inconsistencies into the reconstructed scene, leading to ghosting artifacts, degraded geometry, and reduced visual realism in synthesized views. Consequently, removing transient objects has become increasingly important for applications such as city-scale mapping, infrastructure inspection, construction-site documentation, digital twin creation, and immersive virtual or augmented reality, where temporary objects can permanently corrupt the reconstructed scene and reduce its long-term usability.

Despite recent progress, fully automated transient object removal remains a challenging problem. Existing methods such as T-3DGS (Markin et al., 2024; Sabour et al., 2025; Kulhanek et al., 2024) identify transient regions solely from multi-view geometric inconsistencies, avoiding any dependence on object categories or manually annotated transient masks. While this purely geometry-

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.

driven formulation is broadly applicable, its performance is inherently tied to the quality and consistency of the underlying multi-view reconstruction. In real-world scenes containing irregular camera trajectories, sparse viewpoint coverage, illumination variation, and multiple dynamic objects, transient inference based solely on geometric inconsistency can become unreliable, often producing fragmented or spatially incoherent masks that incorrectly include portions of the static scene. An alternative direction is to leverage the remarkable progress made in GS-based object removal and diffusion-guided inpainting methods, such as InFusion (Liu et al., 2026) and related editing pipelines (Huang et al., 2025; Wang et al., 2024b; 2026; Chen et al., 2024). These approaches can produce highly realistic scene edits once the object to be removed is known. However, they operate under the assumption that accurate object masks are already available before editing begins. While this assumption is reasonable for interactive scene editing (Liu et al., 2026), it becomes a major limitation for automated transient removal, where determining the transient object itself is the primary challenge. A straightforward solution is to detect transient-prone categories, such as people and vehicles, using modern instance segmentation models. Yet semantic detection alone provides no notion of persistence, a detected vehicle may correspond to a passing car that should be removed or a permanently parked vehicle that forms part of the static scene. Consequently, before any editing can be performed, it is necessary to determine not only *what* has been detected, but also *whether it is truly transient*.

Reliable transient authentication alone; however, does not guarantee successful scene editing. Authentication determines *what* should be removed, but it does not recover the scene content that becomes visible once the transient object has been eliminated. A straightforward mechanism is to simply mask the authenticated transient regions and rely on the remaining multi-view observations to recover the missing geometry during Gaussian optimization. In practice, this assumption rarely holds under unconstrained real-world captures, where transient objects frequently occlude large portions of the background and neighboring views provide incomplete or insufficient supervision. Consequently, masking alone often leaves reconstruction voids, degraded textures, and geometric inconsistencies (See supplementary). Recent advances in Gaussian Splatting editing have demonstrated that supplying a generative prior through diffusion-guided image inpainting significantly improves reconstruction quality by synthesizing plausible observations of the occluded background before optimization (Liu et al., 2026; Mirzaei et al., 2023; Wang et al., 2024b).

While diffusion-guided inpainting addresses the missing-content problem, its effectiveness critically depends on *which* images are edited. Existing GS editing pipelines typically perform inpainting on one or a few randomly selected views, implicitly assuming that any edited observation provides sufficient supervision for reconstructing the scene. We argue that this assumption is fundamentally suboptimal for real-world captures. Under sparse or irregular camera trajectories, large viewpoint changes, or severe transient occlusions, a randomly selected view may only partially observe the transient region or in some cases may not provide meaningful supervision for the newly exposed background at all. Even when the transient is clearly visible, a single inpainted image rarely captures enough complementary geometric information to faithfully reconstruct the underlying scene across all viewpoints. To address this limitation, *TRACE* introduces a *geometry-aware minimal spanning camera selection mechanism* that identifies a compact yet maximally informative subset of views by jointly optimizing transient visibility and multi-view geometric coverage. The selected images are subsequently inpainted using a diffusion model and used to optimize the Gaussian representation, enabling consistent propagation of edits throughout the reconstructed scene. Even though introduced here as part of a fully automated transient removal framework, the proposed view selection mechanism is generally applicable to GS based object editing whenever reliable object masks are available.

We extensively evaluate *TRACE* on the Spin-NeRF (Mirzaei et al., 2023), Neural3D (Li et al., 2022a), and WRIVA datasets (Brown et al., 2024), covering both general GS object editing and challenging real-world transient removal scenarios. On Spin-NeRF and Neural3D, we demonstrate that the proposed geometry-aware view selection consistently improves reconstruction quality over existing random view selection strategies. On WRIVA, we further show that coupling transient authentication with geometry-aware editing enables fully automated transient object removal under challenging real-world conditions. See supplementary materials for the related works section.

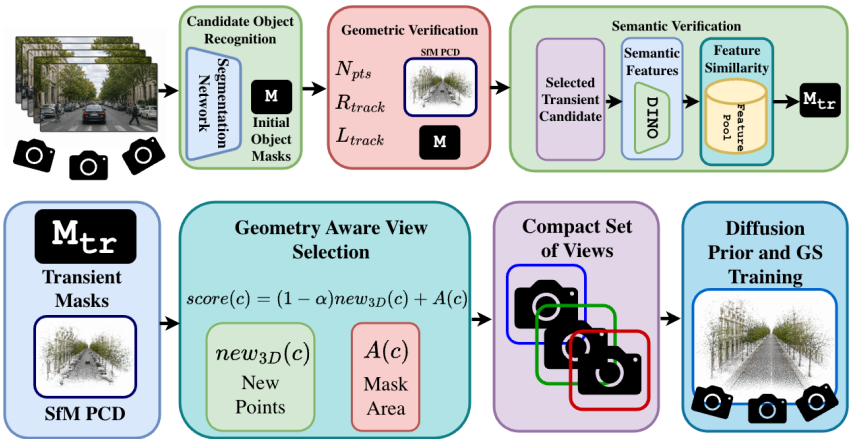


Figure 1: **Overview of the proposed *TRACE* framework:** Given a set of calibrated multi-view images, *TRACE* first identifies candidate transient objects using an instance segmentation network. Candidate masks are subsequently authenticated through multi-view geometric verification using the underlying SfM reconstruction, where object-level geometric support, track consistency, and view coverage provide the primary authentication cues. When geometric evidence is insufficient, semantic verification is performed using DINO features to resolve ambiguous candidates, producing a set of authenticated transient masks $\mathbf{M}_{\text{tr}}$. These verified masks are then used to guide the proposed geometry-aware view selection mechanism, which selects a compact set of informative views by jointly maximizing newly observed 3D geometry and transient visibility. Finally, the selected views are edited using a diffusion prior and employed to optimize a transient-free GS representation. By explicitly separating transient authentication from scene editing, *TRACE* enables reliable and fully automatic transient object removal in challenging real-world multi-view scenes.

2 METHOD

Given a set of multi-view images, *TRACE* first reconstructs the scene using a standard Structure-from-Motion (SfM) pipeline, producing calibrated camera poses and a sparse 3D point cloud that serve as the geometric foundation for the subsequent stages. The goal of *TRACE* is to automatically remove transient objects while preserving the geometric and photometric consistency of the reconstructed Gaussian scene.*TRACE* addresses this problem through a unified pipeline that first determines *what* should be removed and subsequently determines *how* it should be removed (See Figure 1)

The first stage focuses on **transient authentication**, illustrated in the top row of Figure 1. We begin by detecting candidate transient objects belonging to transient-prone semantic categories using an off-the-shelf instance segmentation model. Rather than directly treating these detections as transient, *TRACE* exploits the multi-view geometric information already available from the SfM reconstruction to verify whether each detected object exhibits persistent support across the captured views. Candidate objects that remain geometrically consistent throughout the reconstruction are preserved as part of the static scene, whereas objects lacking sufficient geometric persistence are authenticated as true transient objects. This stage enables fully automated transient identification without requiring manually annotated transient masks.

Once reliable object masks have been obtained, *TRACE* proceeds to **coverage-aware editing**, which is illustrated in the bottom row of Figure 1. Simply masking authenticated transient regions provides insufficient supervision for reconstructing the newly exposed background. Instead, *TRACE* first identifies a compact yet informative subset of views that jointly maximize transient visibility and multi-view geometric coverage. These selected images are subsequently inpainted using a diffusion prior to synthesize plausible background content before Gaussian optimization. Finally, the edited images are used to optimize the Gaussian representation, allowing the reconstructed scene to consistently propagate the edits across all viewpoints while preserving geometric fidelity.

2.1 SCENE REPRESENTATION

Given a set of multi-view images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, we first recover a standard SfM reconstruction consisting of calibrated camera parameters $\mathcal{C} = \{C_i\}_{i=1}^{N}$ and a sparse set of triangulated 3D points $\mathcal{P} = \{\mathbf{p}_k\}_{k=1}^{M}$. Each point $\mathbf{p}_k$ is reconstructed by establishing geometrically consistent correspondences across multiple images and therefore represents a scene element that is supported by one or more observations. For every reconstructed point, SfM additionally provides a visibility track $\mathcal{V}(\mathbf{p}_k) \subseteq \{1, \ldots, N\}$, which records the set of images in which the point is observed. Together, the calibrated cameras, triangulated 3D points, and their visibility tracks establish explicit correspondences between the underlying 3D scene geometry and its image observations. These multi-view geometric relationships provide rich cues about object persistence, visibility, and scene coverage, forming the geometric foundation for the proposed transient authentication and geometry-aware view selection modules.

2.2 TRANSIENT CANDIDATE RECOGNITION

The objective of this stage is to generate a set of candidate objects whose transient status will subsequently be verified. In real-world scenes, transient objects frequently belong to semantic categories such as pedestrians, cars, trucks, buses, motorcycles, bicycles, and other movable objects. Motivated by this observation, ***TRACE*** first performs category-level instance segmentation to identify objects that are likely to exhibit transient behavior. In our implementation, we employ YOLOv11-Seg (Jocher & Qiu, 2024) to obtain a set of instance masks $\mathcal{M}_i = m_i^1, \ldots, m_i^{K_i}$ for each image I_i, together with their corresponding semantic labels and confidence scores. All detections belonging to the predefined transient-prone categories are retained as candidate objects and forwarded to the subsequent authentication stage.

2.3 SFM-GUIDED TRANSIENT AUTHENTICATION

Following candidate object recognition, the next objective is to determine whether each detected object is genuinely transient or forms part of the persistent scene. To this end, ***TRACE*** performs a hierarchical authentication procedure that first analyzes multi-view geometric evidence derived from the underlying SfM reconstruction and subsequently invokes semantic verification only when the available geometric evidence is insufficient.

2.3.1 OBJECT-LEVEL GEOMETRIC ASSOCIATION

The first step of transient authentication is to establish the geometric support associated with each detected object. While candidate recognition operates in the image domain, authentication requires reasoning in the underlying multi-view 3D reconstruction. We therefore associate every candidate object with the subset of SfM points that geometrically support its observations across the reconstructed scene. Let $m_i^j \in \mathcal{M}_i$ denote the j-th candidate object detected in image I_i. Using the calibrated camera parameters C_i, each reconstructed point $\mathbf{p}_k \in \mathcal{P}$ is projected onto the image plane via the projection operator $\Pi_{C_i}(\cdot)$. The object-specific point set is then defined as $\mathcal{P}(m_i^j) = \left\{ \mathbf{p}_k \in \mathcal{P} \;\middle|\; \Pi_{C_i}(\mathbf{p}_k) \in m_i^j \right\}$, where $\Pi_{C_i}(\mathbf{p}_k) \in m_i^j$ indicates that the projected point lies inside the segmentation mask. Consequently, $\mathcal{P}(m_i^j)$ contains all reconstructed 3D points geometrically associated with the candidate object. Unlike image-space detections, the associated SfM points provide access to rich multi-view geometric information. Collectively, these visibility tracks establish the multi-view support of the detected object, enabling reasoning about its geometric persistence across the entire image collection rather than within a single image. The resulting object-level geometric representation serves as the foundation for the persistence analysis described in the following subsection.

2.3.2 MULTI-VIEW PERSISTENCE VERIFICATION

Once the object-level geometric support has been established, ***TRACE*** quantifies the persistence of each candidate object by analyzing the multi-view properties of its associated SfM points. Rather than relying on a single geometric cue, we characterize persistence through three complementary measurements that capture the *strength, consistency, and spatial distribution* of the reconstructed

geometric evidence. Together, these cues provide a robust description of whether a detected object belongs to the underlying static scene or corresponds to a genuine transient observation.

Geometric Point Support: Persistent objects are typically reconstructed from a larger number of stable feature correspondences than transient objects. We therefore first measure the amount of reconstructed geometry supporting a candidate by $N_{\text{pts}}(m_i^j) = |\mathcal{P}(m_i^j)|$, where $\mathcal{P}(m_i^j)$ denotes the set of SfM points associated with the candidate object. Objects with limited geometric support generally indicate insufficient multi-view evidence and are more likely to correspond to transient observations.

Track Consistency: The existence of reconstructed points alone does not necessarily imply object persistence. We therefore analyze how consistently these points are observed across the image collection. For each associated point $\mathbf{p}_k \in \mathcal{P}(m_i^j)$, its visibility track, $\mathcal{V}(\mathbf{p}_k)$ records the set of images in which the point is successfully reconstructed. The average track support of the candidate is computed as $L_{\text{track}}(m_i^j) = \frac{1}{|\mathcal{P}(m_i^j)|} \sum_{\mathbf{p}_k \in \mathcal{P}(m_i^j)} |\mathcal{V}(\mathbf{p}_k)|$. Large values of L_{track} indicate that the associated geometry is repeatedly observed from multiple viewpoints, whereas transient objects typically exhibit shorter and less consistent visibility tracks.

View Coverage: Eventually, we measure how broadly the geometric support is distributed across the image collection. Let $\mathcal{U}(m_i^j) = \bigcup_{\mathbf{p}_k \in \mathcal{P}(m_i^j)} \mathcal{V}(\mathbf{p}_k)$ denote the set of unique images that observe the candidate geometry. We define the view coverage ratio as $R_{\text{view}}(m_i^j) = \frac{|\mathcal{U}(m_i^j)|}{N}$, where N is the total number of input images. Unlike the average track length, this quantity measures how widely the object is supported throughout the complete capture rather than how many observations each individual point receives.

Taken together, N_{pts}, L_{track}, and R_{view} describe complementary aspects of geometric persistence. Persistent scene objects generally exhibit strong geometric support, long and consistent visibility tracks, and broad view coverage, whereas genuine transient objects tend to receive substantially weaker support across one or more of these dimensions. The resulting persistence measurements form the primary evidence used by *TRACE* for transient authentication.

2.3.3 Semantic Verification under Weak Geometry

The geometric persistence cues described above provide reliable authentication for the majority of detected objects. However, SfM reconstruction quality may degrade in challenging scenarios, including distant objects, motion blur, textureless surfaces, partial occlusions, or sparse viewpoint coverage. In these cases, the associated geometric support may be insufficient to confidently determine whether a candidate belongs to the static scene or represents a genuine transient object. Rather than forcing a decision from incomplete geometric evidence, *TRACE* performs a secondary semantic verification step for geometrically ambiguous candidates. Specifically, we compare the visual representation of the candidate object against a pool of authenticated persistent objects belonging to the same semantic category. Let $\phi(\cdot)$ denote the image embedding extracted from a pre-trained vision encoder (CLIP (Radford et al., 2021) or DINOv2 (Oquab et al., 2023)). For an ambiguous candidate m_i^j, its semantic consistency is computed as $S_{\text{sem}}(m_i^j) = \max_{m_p \in \mathcal{S}_c} \cos\Big(\phi(m_i^j), \phi(m_p)\Big)$, where $\mathcal{S}_c$ denotes the set of authenticated persistent objects belonging to semantic category c, m_p is an authenticated persistent object in $\mathcal{S}_c$, $\phi(\cdot)$ denotes the DINO feature representation, and $\cos(\cdot,\cdot)$ is the cosine similarity. A high semantic consistency indicates that the candidate exhibits appearance characteristics similar to objects that have already been verified as persistent, providing additional evidence that it should be retained as part of the static scene. Conversely, candidates that remain unsupported by both geometric persistence and semantic consistency are authenticated as transient objects. Importantly, semantic verification is invoked only when geometric evidence is insufficient, ensuring that the authentication process remains fundamentally geometry-driven while maintaining robustness under challenging reconstruction conditions.

2.4 Coverage-aware View Selection

Following transient authentication, *TRACE* obtains a set of verified transient masks $\mathcal{M}_{\text{tr}}$. A straightforward mechanism is to simply mask the transient regions and optimize the Gaussian representation

using the remaining observations. However, masking alone leaves the newly exposed background unsupervised, often resulting in reconstruction voids and degraded geometry. Recent GS editing methods therefore employ diffusion-guided image inpainting to synthesize the missing scene content prior to Gaussian optimization (Liu et al., 2026; Mirzaei et al., 2023). The effectiveness of this mechanism, however, depends critically on *which* views are selected for editing. Existing GS based object-removal/inpainting pipelines typically inpaint one or a few randomly selected views. Such a mechanism does not account for the underlying multi-view geometry and is unlikely to provide sufficient supervision under sparse camera trajectories, large viewpoint changes, or severe object occlusions. A selected image may only partially observe the transient, while multiple selected views may repeatedly capture nearly identical scene content, providing little additional geometric information. Consequently, the edited observations may fail to consistently supervise the reconstruction of the newly revealed background. To address this limitation, ***TRACE*** formulates view selection as an object-aware geometric coverage problem. Given the calibrated cameras $\mathcal{C}$, reconstructed SfM points $\mathcal{P}$, and authenticated transient masks $\mathcal{M}_{\text{tr}}$, our objective is to identify a compact subset of cameras $\mathcal{C}^* \subseteq \mathcal{C}$ that jointly: 1). Maximize the geometric coverage of the reconstructed scene,2). Maximize the visibility of the authenticated transient regions, 3). Minimize the number of images requiring diffusion-guided inpainting.

For every camera $c \in \mathcal{C}$, the SfM reconstruction naturally provides the set of visible 3D points $\mathcal{V}(c)$. Using the authenticated transient mask, we further identify the subset of visible points belonging to the transient object $\mathcal{O}(c) = \{p \in \mathcal{V}(c) \mid \Pi_c(p) \in \mathcal{M}_{\text{tr}}\}$, where $\Pi_c(\cdot)$ denotes the camera projection. The object mask area $A(c)$ provides an additional measure of how prominently the transient is observed in the corresponding image. We evaluate each candidate camera according to the amount of *previously uncovered* scene geometry it contributes together with its object visibility, $\text{score}(c) = (1-\alpha)\,\text{new}_{\text{3D}}(c) + \alpha\,A(c)$,, where $\text{new}_{\text{3D}}(c)$ denotes the number of reconstructed points newly covered by camera c beyond those already explained by the selected views, and α is a hyperparameter. The camera with the largest marginal gain is iteratively selected until the desired budget is reached or no additional geometric coverage can be obtained. The complete greedy optimization procedure is summarized in Algorithm 1 in the supplementary material. The selected camera set $\mathcal{C}^*$ is subsequently inpainted using a diffusion model (Yu et al., 2023), and the edited images are used to optimize the Gaussian Splatting representation. While *TRACE* employs the authenticated transient masks $\mathcal{M}_{\text{tr}}$ to enable fully automatic transient removal, the proposed geometry-aware view selection strategy is independent of how the object masks are obtained. Consequently, it can be seamlessly integrated with existing Gaussian Splatting-based object editing and inpainting pipelines, providing stronger geometric supervision for diffusion-guided editing and more consistent scene reconstruction than conventional random view selection.

3 Experiments

3.1 Evaluation of Transient Authentication

We first evaluate the proposed transient authentication module independently of the subsequent diffusion-guided editing stage. The objective of this experiment is to assess whether the method can reliably identify transient objects under challenging real-world capture conditions. We consider two representative WRIVA Challenge sequences. The first, named as, `t01_v01_s04_r08_ImageDensity_A01_iPad_LTS`, captures a building entrance and parking area containing numerous transient objects, including pedestrians, and moving or temporarily parked vehicles. The scene is acquired using highly irregular camera trajectories with substantial viewpoint variation, making transient identification considerably more difficult than in controlled multi-view datasets (See the supplementary material for a comparison of camera trajectories in Spin-NeRF, Neural-3D, and WRIVA datasets). The second sequence, named as, `t05_v09_s00_r06_ArtifactInjection_M04_car`, contains a permanently broken vehicle that forms part of the static environment. This sequence evaluates whether the method can distinguish persistent objects from genuinely transient instances of the same semantic category. *TRACE* first detects candidate transient-prone objects using instance segmentation and subsequently au-

Table 1: **Quantitative comparison on two representative WRIVA Challenge sequences.**

	Dataset_1		Dataset_2	
Method	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
T-3DGS	17.87	0.4813	**30.0223**	0.8799
TRACE	**19.37**	**0.5099**	29.8029	**0.8876**

thenticates each candidate through the proposed SfM-guided multi-view persistence verification. The resulting authenticated transient masks are then used during Gaussian Splatting optimization by excluding the masked regions from the reconstruction loss. No diffusion-guided inpainting is employed in this experiment, allowing us to isolate the effectiveness of transient authentication alone. The comparison method is the recently proposed T-3DGS, which directly infers transient regions from geometric inconsistencies without semantic verification or persistence authentication. Figure 2 presents representative qualitative comparisons. On the transient-heavy `ImageDensity` sequence, *TRACE* successfully authenticates the pedestrian and produces a complete transient mask, leading to a clean reconstruction in which the pedestrian is removed despite the absence of diffusion-guided editing. In contrast, the transient masks produced by T-3DGS are highly fragmented and fail to localize the pedestrian, resulting in the person remaining visible in the rendered scene. This failure illustrates the limitations of relying solely on geometric inconsistency under sparse viewpoints and irregular camera motion.

The second WRIVA sequence further evaluates false-positive behavior on a persistent object. Even though the scene contains a vehicle, the object is a permanent component of the environment rather than a transient. *TRACE* correctly authenticates the vehicle as persistent and therefore excludes it from the transient masks. Additional qualitative examples for both sequences are provided in the supplementary material. Table 1 reports the corresponding quantitative results. As can be seen *TRACE* consistently achieves higher reconstruction structural quality on transient scenes while avoiding the erroneous removal of persistent scene content. These results demonstrate that reliable transient removal fundamentally depends on accurate transient authentication before any editing is performed.

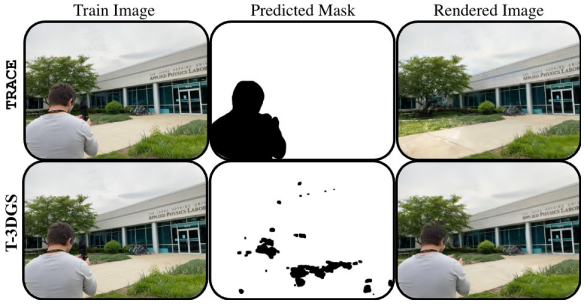


Figure 2: **Evaluation of transient authentication on the WRIVA dataset.** *TRACE* produces an accurate transient mask, enabling successful pedestrian removal through masked Gaussian optimization. In contrast, T-3DGS produces fragmented, spatially incoherent masks under irregular viewpoints, leaving the pedestrian visible and degrading other regions.

While *TRACE* is designed as a fully automatic framework for transient object removal, its second contribution, which is, coverage-aware view selection, is independent of transient authentication. Once an object mask is available, regardless of how it is obtained, the proposed view selection mechanism can be directly integrated into existing GS editing pipelines. Consequently, we additionally evaluate *TRACE* on the Spin-NeRF dataset, which contains manually specified object removal tasks rather than transient objects. This experiment isolates the proposed view selection module and demonstrates that its benefits extend beyond transient removal to general 3D scene editing. Unlike WRIVA, where transient objects must first be identified and authenticated, Spin-NeRF assumes the object masks are given (or can easily be obtained). Existing diffusion-guided editing methods, such as InFusion and GScream, typically perform inpainting on randomly selected training views before Gaussian optimization. We argue that random selection does not necessarily provide the strongest

Scene	InFusion	+*TRACE*	GScream	+*TRACE*
2	17.33	**17.52**	18.79	**18.91**
4	16.87	**20.60**	20.36	**22.69**
7	20.04	**20.94**	21.27	**21.43**
10	18.37	**18.76**	19.78	**19.84**
Book	22.17	**22.17**	23.57	**23.68**
Trash	22.03	**22.17**	23.39	**24.29**

Table 2: **Quantitative results on Spin-NeRF.** PSNR comparison of InFusion and GScream with and without *TRACE*. Higher is better.

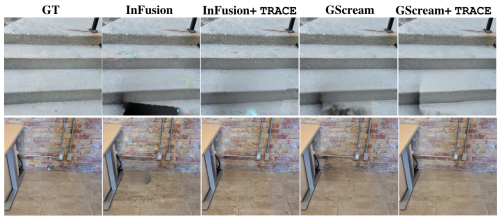
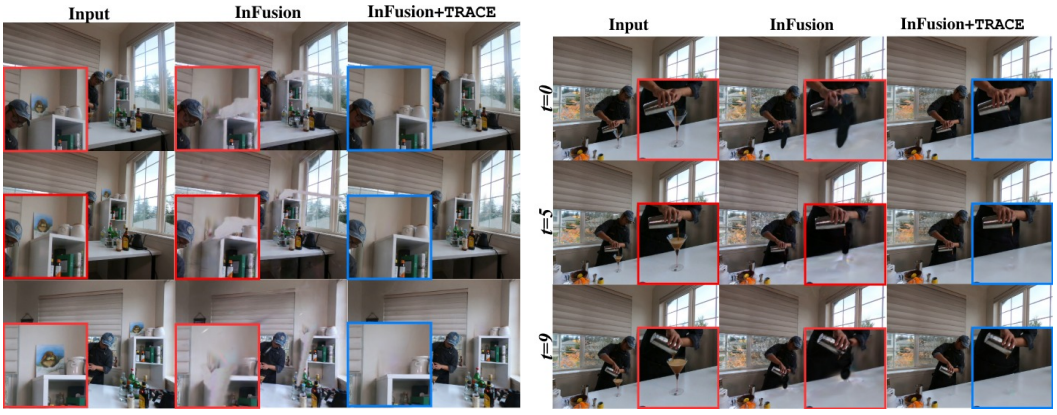


Figure 3: **Qualitative results on Spin-NeRF.** Integrating *TRACE* into InFusion and GScream yields more coherent object removal and sharper reconstruction of the newly exposed background.



(a) *Calendar* scene. (b) *Coffee_martini* scene.

Figure 4: **Object removal on Neural 3D scenes.** *TRACE* replaces random view selection in InFusion with coverage-aware selection, producing cleaner object removal and more consistent reconstruction of the newly exposed background across views and timestamps.

supervision for reconstruction, as the selected views may only partially observe the object or contain highly redundant geometric information. Instead, *TRACE* identifies a compact set of complementary views that jointly maximize both geometric coverage and object visibility, providing more informative supervision for diffusion-guided inpainting. Table 2 reports quantitative comparisons after integrating the proposed view selection strategy into both InFusion and GScream. Across representative Spin-NeRF scenes, *TRACE* consistently improves reconstruction quality over the original pipelines, demonstrating that selecting geometrically informative views is more effective than random view selection. Figure 3 further illustrates the qualitative improvements. Compared with the original editing pipelines, *TRACE* produces cleaner object removal, sharper reconstruction of newly exposed background regions, and more view-consistent geometry. These results demonstrate that the proposed coverage-aware view selection strategy is not limited to transient removal, but constitutes a general plug-and-play enhancement for diffusion-guided Gaussian Splatting editing frameworks.

3.2 Generalization to Sparse Multi-view Scenes: Neural 3D

To further evaluate the proposed coverage-aware view selection strategy, we consider the Neural 3D dataset (Research, 2024), which presents a substantially more challenging setting than Spin-NeRF. Unlike the dense and smoothly varying camera trajectories of Spin-NeRF, Neural 3D exhibits sparse and irregular viewpoint distributions, where objects are often observed by only a limited subset of cameras. Under such conditions, random view selection frequently fails to provide sufficient supervision for diffusion-guided inpainting, making view selection considerably more critical.

We integrate *TRACE* into recently proposed InFusion without modifying its inpainting or optimization pipeline. Qualitative results on the *coffee_martini* scene are shown in Figure 4b, where the task is to remove the martini glass across multiple timestamps. Figure 4a presents an additional object removal example involving a picture placed on top of a cupboard. In both cases, *TRACE* consistently produces more coherent object removal and sharper reconstruction of the newly exposed background than the original pipeline.As Neural 3D does not provide ground-truth images for edited scenes, we adopt an object re-detection protocol to objectively assess removal quality. Following object removal, a YOLOv8x detector is applied to the rendered views, and any remaining detections within the edited regions are treated as residual object evidence. Table 3 summarizes the results on two representative timestamps of the *coffee_martini* sequence. Compared with the original InFusion pipeline, *TRACE* completely suppresses

Table 3: **Residual object detection on the Neural 3D dataset.** Lower values indicate more complete object removal.

Time	Method	Detections	Mean IoU↓
t_1	InFusion	4/17	0.0744
	TRACE	**0/17**	**0.0000**
t_9	InFusion	0/17	0.0000
	TRACE	**0/17**	**0.0000**

residual martini-glass detections, demonstrating that selecting geometrically informative views provides more effective supervision for diffusion-guided inpainting under sparse multi-view capture.

3.3 Results on the WRIVA Dataset

The WRIVA dataset represents the primary evaluation benchmark for the complete *TRACE* framework. Unlike Spin-NeRF and Neural 3D, WRIVA contains real-world scenes captured under highly irregular camera trajectories, significant viewpoint variation, and numerous transient objects, making both transient identification and subsequent object removal substantially more challenging. These characteristics closely reflect the conditions encountered in practical large-scale scene reconstruction, where transient objects are neither known a priori nor consistently observed across views. For transient identification, *TRACE* employs the proposed SfM-guided authentication module to automatically distinguish genuine transient objects from persistent scene content.



Figure 5: **Transient removal and novel-view rendering on a challenging WRIVA sequence.** The black jeep is never fully visible from a single view, while irregular camera motion and large viewpoint changes make transient authentication and multi-view removal particularly challenging. Despite these conditions, the reconstructed scene supports consistent novel-view rendering after transient removal.

Rather than comparing directly with geometry-based transient localization methods such as T-3DGS, we evaluate the downstream object removal stage using recent diffusion-guided Gaussian Splatting editing frameworks. This provides a fair assessment of editing quality, as T-3DGS primarily targets transient localization and often produces fragmented or spatially unstable masks under challenging real-world conditions. In contrast, *TRACE* provides geometrically authenticated transient masks, enabling an isolated evaluation of the proposed coverage-aware editing strategy. We evaluate the proposed framework on the `t01_v01_s04_r08_ImageDensity_A01_iPad_LTS` sequence. Representative observations of the front side of the building are shown in Figure 5 (top). This dataset is particularly challenging because the transient objects are never fully visible from a single view, the camera motion is highly irregular, and numerous additional transient objects appear throughout the scene. Figure 5 (bottom) presents qualitative comparisons after integrating *TRACE* into the diffusion-guided editing pipeline, viewed from a novel viewpoint. Compared with the InFusion, *TRACE* produces substantially cleaner transient removal, preserves road markings and surrounding textures, and avoids the streaking and geometric distortions introduced by random view selection. These results demonstrate the effectiveness of combining authenticated transient identification with coverage-aware multi-view editing for real-world transient object removal.

4 Additional Results

The supplementary material includes an extended discussion of related work, additional evaluation results, training configurations, hyperparameters, and both quantitative and qualitative analyses.

5 Conclusion

We introduced TRACE, a fully automatic framework for transient object removal in Gaussian Splatting that authenticates candidate objects before editing and leverages geometry-aware view selection for efficient multi-view inpainting. Extensive experiments show that TRACE consistently improves transient localization and reconstruction quality across diverse real-world scenes. More broadly, our results highlight the importance of reasoning about object persistence prior to scene editing and suggest a promising direction for robust, geometry-aware 3D scene reconstruction.

6 AI USE STATEMENT

In this work, we used generative AI tools for grammar correction and writing enhancement, as well as to assist in generating some methodological figures. All AI-assisted text and figures were carefully reviewed and verified by the authors for accuracy and consistency with the proposed methodology. Generative AI was not used for research ideation, methodology development, data generation or analysis, experimental design, or the formulation of scientific claims and conclusions. We take full responsibility for the final content of this work, including text and figures produced with the assistance of generative AI.

REFERENCES

Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001.

M Bertalmio. Image inpainting, 2000.

Myron Brown, Michael Chan, and Michael Twardowski. Wriva public data, 2024. URL `https://dx.doi.org/10.21227/cjk5-gf33`.

Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024.

Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21476–21485, 2024.

Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 811–820, 2024.

Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. In *Computer Graphics Forum*, volume 31, pp. 305–314. Wiley Online Library, 2012.

Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14304–14314. IEEE Computer Society, 2021.

Sankeerth Durvasula, Adrian Zhao, Fan Chen, Ruofan Liang, Pawan Kumar Sanjaya, and Nandita Vijaykumar. Distwar: Fast differentiable rendering on raster-based rendering pipelines. *arXiv preprint arXiv:2401.05345*, 2023.

Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pp. 1033–1038. IEEE, 1999.

Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Sharath Girish, Kamal Gupta, and Abhinav Shrivastava. Eagles: Efficient accelerated 3d gaussians with lightweight encodings. *arXiv preprint arXiv:2312.04564*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Miguel Granados, Kwang In Kim, James Tompkin, Jan Kautz, and Christian Theobalt. Background inpainting for videos with dynamic objects and a free-moving camera. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pp. 682–695. Springer, 2012.

Junfu Guo, Yu Xin, Gaoyi Liu, Kai Xu, Ligang Liu, and Ruizhen Hu. Articulatedgs: Self-supervised digital twin modeling of articulated objects using 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27144–27153, 2025.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 634–644, 2024.

Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G Schwing. Proposal-based video completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pp. 38–54. Springer, 2020.

Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014.

Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (ToG)*, 35(6):1–11, 2016.

Sheng-Yu Huang, Zi-Ting Chou, and Yu-Chiang Frank Wang. 3d gaussian inpainting with depth-guided cross-view consistency. In *Proceedings of the computer vision and pattern recognition conference*, pp. 26704–26713, 2025.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024.

Han Jiang, Haosen Sun, Ruoxuan Li, Chi-Keung Tang, and Yu-Wing Tai. Inpaint4dnerf: Promptable spatio-temporal nerf inpainting with generative diffusion models. *arXiv preprint arXiv:2401.00208*, 2023.

Joongho Jo, Hyeongwon Kim, and Jongsun Park. Identifying unnecessary 3d gaussians using clustering for fast rendering of 3d gaussian splatting. *arXiv preprint arXiv:2402.13827*, 2024.

Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. URL `https://github.com/ultralytics/ultralytics`.

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.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 505–515, 2024.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv preprint arXiv:2407.08447*, 2024.

Junseo Lee, Seokwon Lee, Jungi Lee, Junyong Park, and Jaewoong Sim. Gscore: Efficient radiance field rendering via architectural support for 3d gaussian splatting. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pp. 497–511, 2024.

Seongmin Lee, Benjamin Hoover, Hendrik Strobelt, Zijie J Wang, ShengYun Peng, Austin Wright, Kevin Li, Haekyu Park, Haoyang Yang, and Duen Horng Chau. Diffusion explainer: Visual explanation for text-to-image stable diffusion. *arXiv preprint arXiv:2305.03509*, 2023.

Deqi Li, Shi-Sheng Huang, Zhiyuan Lu, Xinran Duan, and Hua Huang. St-4dgs: Spatial-temporally consistent 4d gaussian splatting for efficient dynamic scene rendering. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024a.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20775–20785, 2024b.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5521–5531, 2022a.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17562–17571, 2022b.

Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5166–5175, 2024.

Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Ping Luo, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5726–5738, 2026.

Haimin Luo, Min Ouyang, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Gaussianhair: Hair modeling and rendering with light-aware gaussians. *arXiv preprint arXiv:2402.10483*, 2024.

Alexander Markin, Vadim Pryadilshchikov, Artem Komarichev, Ruslan Rakhimov, Peter Wonka, and Evgeny Burnaev. T-3dgs: Removing transient objects for 3d scene reconstruction. *arXiv preprint arXiv:2412.00155*, 2024.

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.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20669–20679, 2023.

Hammer Missions. Comparing nerf, 3d gaussian splatting, and drone photogrammetry in geospatial applications, February 2023. URL `https://www.hammermissions.com/post/comparing-nerf-3d-gaussian-splatting-and-drone-photogrammetry`. Accessed: 2023-11-11.

Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 788–798, 2024.

KL Navaneet, Kossar Pourahmadi Meibodi, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Compact3d: Compressing gaussian splat radiance field models with vector quantization. *arXiv preprint arXiv:2311.18159*, 2023.

Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10349–10358, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Avinash Paliwal, Xilong Zhou, Wei Ye, Jinhui Xiong, Rakesh Ranjan, and Nima Khademi Kalantari. Ri3d: Few-shot gaussian splatting with repair and inpainting diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25094–25103, 2025.

Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*, pp. 111–149. Springer, 2021.

AKM Rabby and Chengcui Zhang. Beyondpixels: A comprehensive review of the evolution of neural radiance fields. *arXiv preprint arXiv:2306.03000*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sameera Ramasinghe and Simon Lucey. Beyond periodicity: Towards a unifying framework for activations in coordinate-mlps. In *European Conference on Computer Vision*, pp. 142–158. Springer, 2022.

Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. Dlformer: Discrete latent transformer for video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3511–3520, 2022.

Facebook Research. Neural 3d video. `https://github.com/facebookresearch/Neural_3D_Video/releases/tag/v1.0`, 2024. URL `https://github.com/facebookresearch/Neural_3D_Video/releases/tag/v1.0`. Version 1.0.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David Fleet, and Andrea Tagliasacchi. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *ACM Transactions on Graphics*, 44(2):1–11, 2025.

Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18507–18516, 2023.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 175–184, 2019.

David Tschumperlé and Rachid Deriche. Vector-valued image regularization with pdes: A common framework for different applications. *IEEE transactions on pattern analysis and machine intelligence*, 27(4):506–517, 2005.

Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2626–2634, 2017.

Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 836–850. Springer, 2014.

Cairong Wang, Yiming Zhu, and Chun Yuan. Diverse image inpainting with normalizing flow. In *European conference on computer vision*, pp. 53–69. Springer, 2022.

Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20902–20911, 2024a.

Shaoxiang Wang, Shihong Zhang, Christen Millerdurai, Rüdiger Westermann, Didier Stricker, and Alain Pagani. Inpaint360gs: Efficient object-aware 3d inpainting via gaussian splatting for 360deg scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 117–127, 2026.

Yuxin Wang, Qianyi Wu, Guofeng Zhang, and Dan Xu. Learning 3d geometry and feature consistent gaussian splatting for object removal. In *European conference on computer vision*, pp. 1–17. Springer, 2024b.

Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time completion of video. *IEEE Transactions on pattern analysis and machine intelligence*, 29(3):463–476, 2007.

Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. Surface light fields for 3d photography. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 487–496. 2023.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024a.

Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. In *European conference on computer vision*, pp. 55–71. Springer, 2024b.

Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2024.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European conference on computer vision*, pp. 162–179. Springer, 2024.

Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 69–78, 2021.

Hongjia Zhai, Xiyu Zhang, Boming Zhao, Hai Li, Yijia He, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Splatloc: 3d gaussian splatting-based visual localization for augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Cong Zhang, Wenxia Yang, Xin Li, and Huan Han. Mmginpainting: Multi-modality guided image inpainting based on diffusion models. *IEEE Transactions on Multimedia*, 2024.

Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5741–5750, 2020.

Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *European Conference on Computer Vision*, pp. 277–296. Springer, 2022.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *European Conference on Computer Vision*, pp. 145–163. Springer, 2025.

Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16448–16457, 2021.

A Appendix

A.1 Related Works

A.1.1 3D/4D representations

Novel view synthesis aims to generate realistic images from new viewpoints, a core challenge in computer vision and graphics. While techniques like light field interpolation (Davis et al., 2012) achieve high-quality results with dense viewpoints, sparse-view scenarios often rely on mesh-based representations to model scene surfaces with diffuse or view-dependent properties (Waechter et al., 2014; Wood et al., 2023). However, these methods frequently struggle with optimization due to poor conditioning in the loss landscape (Mildenhall et al., 2021). Volumetric representations are another approach to novel view rendering, generating photorealistic images from RGB inputs. Early methods used voxel grids, later enhanced by deep neural networks (Srinivasan et al., 2019; Tulsiani et al., 2017; Zhou et al., 2018). NeRF (Mildenhall et al., 2021) transformed the field by encoding scenes with neural networks, leveraging specific activation functions (Ramasinghe & Lucey, 2022; Saragadam et al., 2023) or embedding layers. While advances extend to 3D and 4D representations (Du et al., 2021), NeRF's implicit nature limits real-time performance and editability (Rabby & Zhang, 2023; Missions, 2023). As a solution, 3DGS (Kerbl et al., 2023) has been introduced, a method that leverages the sparse point clouds from SfM. Unlike NeRF-based techniques. Significant advancements have been made to 3DGS, by focusing on improving storage efficiency (Navaneet et al., 2023; Girish et al., 2023; Niedermayr et al., 2024) , training efficiency (Durvasula et al., 2023), rendering speed (Jo et al., 2024; Lee et al., 2024), generalization for handling, sparse views (Chung et al., 2024; Zhu et al., 2025; Li et al., 2024b) and "in-the-wild" scenarios (Kulhanek et al., 2024). Alongside these technical improvements, numerous applications have also been emerged, including human reconstruction (Moreau et al., 2024; Kocabas et al., 2024; Hu et al., 2024) and hair (Xu et al., 2024) and head modeling (Luo et al., 2024). 3DGS has also been extended to the fourth dimension, which is time, where a 3DGS model is trained for each timestamp for a certain number of epochs, thereafter, it considers the inherent continuity of the objects motion, and is maintained by predicting Gaussian deformation with a multi-head Gaussian deformation decoder. The deformed Gaussians are then splatted to produce rendered images, enhancing dynamic scene representation (Wu et al., 2024a; Li et al., 2024a).

A.1.2 Image, and Video Inpainting

Image and video inpainting aim to fill masked regions with content that blends seamlessly with the background. Early image inpainting relied on Partial Differential Equations (PDE) (Bertalmio, 2000; Ballester et al., 2001; Tschumperlé & Deriche, 2005) and patch-based methods (Efros & Leung, 1999; Huang et al., 2014). When it comes to video inpainting, it is more challenging due to the need for temporal consistency, with early methods using patch sampling and synthesis (Wexler et al., 2007; Granados et al., 2012; Huang et al., 2016). On the other hand, Deep learning (DL) models leverage large datasets to learn patterns, textures, and dynamics. DL-based inpainting includes deterministic and stochastic methods, with the latter using VAEs (Pinheiro Cinelli et al., 2021; Zhao et al., 2020), GANs (Goodfellow et al., 2020; Zheng et al., 2022), flow-based models (Rezende & Mohamed, 2015; Wang et al., 2022), MLMs (Yu et al., 2021), and diffusion models (Zhang et al., 2024; Huang et al., 2024). DL based video inpainting techniques mainly fall into four categories: 3D CNNs (Hu et al., 2020), shift-based methods (Zou et al., 2021), flow-guided techniques (Li et al., 2022b), and attention-based models (Ren et al., 2022), each offering unique advantages for spatial and temporal consistency.

A.1.3 3D Scene Editing

With advances in 3D computer vision and graphics, there has been a significant growth in the editability of learned 3D and 4D representations. As outlined in section 1 , implicit representations encode scenes within network weights and biases, requiring specialized algorithms to enable scene editing. For example, a way to edit the scene is through introducing perceptual optimization approach that leverages 2D inpainting models to transfer information into 3D space while ensuring view consistency (Mirzaei et al., 2023). The 4D extension of NeRF inpainting has been explored by iteratively applying Stable Diffusion to inpaint and render training images through NeRF, achieving

object removal and multi-view consistency across frames (Lee et al., 2023), (Jiang et al., 2023). When it comes to the GS domain, editability typically involves sequentially editing multi-view 2D images before reconstructing the 3D Gaussian representation. The approach proposed in (Wang et al., 2024a) utilizes multi-modal models, including large language and segmentation models, to predict editable regions from textual descriptions. While scene edits like object recoloring require special attention, inpainting with 3DGS demands even more careful handling to ensure consistency across views. More recently, A 3DGS inpainting mechanism was introduced in (Liu et al., 2026), which performs image-conditioned depth inpainting from a single inpainted camera view. However, challenges remain, most notably the tendency of the trained Gaussians to overfit the selected view and the difficulty of maintaining consistency across multiple viewpoints. To further enhance information propagation from the inpainted 2D view to the masked region, an attention-based mechanism has also been proposed (Wang et al., 2024b). More recently, a GS inpainting method aimed at preserving cross-view consistency through depth regularization was introduced in (Huang et al., 2025). In addition to static object removal, T-3DGS (Markin et al., 2024) effectively removes transient objects from 3D reconstructions. A further notable research efforts on this direction includes (Wang et al., 2026; Paliwal et al., 2025; Huang et al., 2025). As demonstrated throughout our experiments in the main paper, the proposed *TRACE* framework extends beyond transient object removal. In particular, the proposed coverage-aware editing module is model-agnostic and can be seamlessly integrated into existing Gaussian Splatting-based object removal and inpainting frameworks that rely on 2D inpainted view supervision. By replacing conventional random view selection with geometry-aware view selection, *TRACE* consistently provides stronger multi-view supervision, leading to more accurate, view-consistent, and geometrically faithful reconstructions.

ADDITIONAL EXPLANATIONS AND EXPERIMENTS

A.2 TRANSIENT AUTHENTICATION DETAILS

As our transient authentication experiments are conducted on the challenging WRIVA datasets, we employ a single set of authentication parameters across all WRIVA Challenge sequences without any scene-specific parameter tuning (Brown et al., 2024). Candidate transient objects are first detected using an off-the-shelf instance segmentation model and subsequently verified through the proposed hierarchical SfM-guided authentication framework.

Geometric Authentication: The implementation follows the hierarchical authentication mechanism described in the main paper. The primary authentication stage relies entirely on multi-view geometric evidence derived from the underlying SfM reconstruction. Specifically, the three persistence measurements introduced in Section 2 are converted into deterministic authentication criteria using fixed geometric thresholds. For each detected object, *TRACE* evaluates

- Geometric point support,
- Feature-track consistency, and
- Cross-view coverage,

through the corresponding quantities N_{pts}, L_{track}, and R_{view}. Persistent objects exhibit strong geometric support, long visibility tracks, and broad multi-view coverage. During implementation, a candidate is immediately authenticated as persistent only when its associated geometry satisfies the predefined criteria. Across all WRIVA Challenge datasets, we employ a single set of fixed thresholds requiring approximately 4–5 feature-track observations, 10–14 supporting triangulated 3D points, and consistent visibility across 7–10 unique views. No scene-specific parameter tuning is performed.

Mask Refinement: To improve the quality of the generated transient masks, we additionally apply lightweight morphological refinement consisting of a small erosion followed by dilation, together with shadow association for nearby vehicle shadows. Unless otherwise stated, every parameter remains fixed across all WRIVA Challenge datasets, demonstrating that *TRACE* generalizes without dataset-specific threshold tuning.

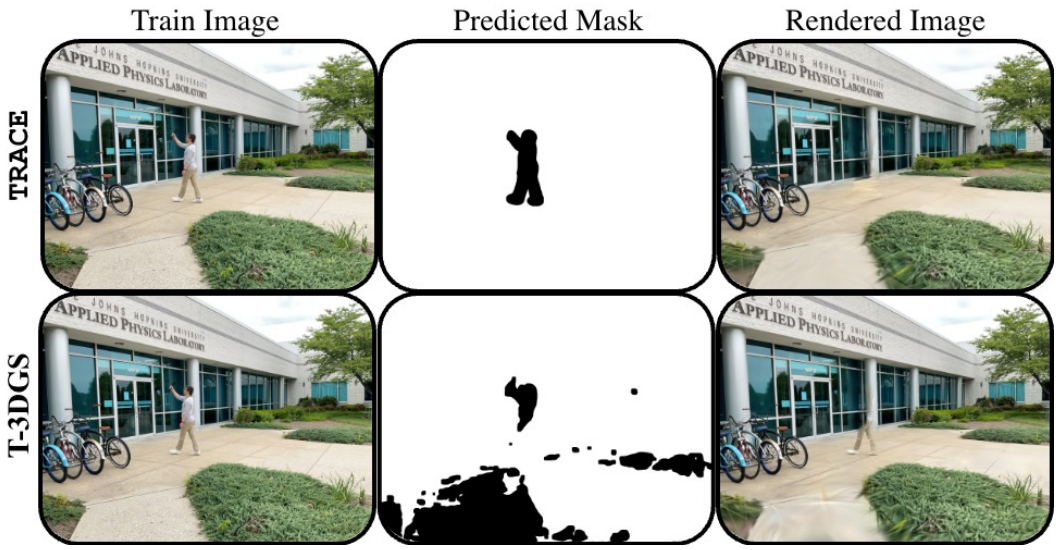


Figure 6: **Transient authentication on the WRIVA `t01_v01_s04_r08_ImageDensity_A01_iPad_LTS` sequence.** *TRACE* combines candidate object recognition with the proposed SfM-guided transient authentication to accurately identify the pedestrian as a transient object, producing a complete transient mask that enables successful removal during Gaussian Splatting optimization. In contrast, T-3DGS relies solely on geometric inconsistency, resulting in fragmented masks that both fail to localize the pedestrian and incorrectly classify portions of the static pavement as transient. Consequently, the pedestrian remains visible in the rendered reconstruction, accompanied by noticeable blur and geometric artifacts.

A.2.1 Additional Transient Authentication Results

Figure 6 highlights a representative failure case of T-3DGS on the WRIVA `ImageDensity` sequence. Owing to highly irregular camera trajectories, sparse viewpoint overlap, and multiple transient objects, geometry alone becomes insufficient for reliable transient localization. As a result, T-3DGS produces fragmented transient masks that simultaneously miss the pedestrian and incorrectly classify portions of the static pavement as transient. These localization errors directly propagate to the reconstructed scene, leaving the pedestrian visible while introducing noticeable blur and geometric distortions in the rendered image. In contrast, *TRACE* authenticates the pedestrian through multi-view geometric persistence, producing an accurate transient mask and a substantially cleaner reconstruction.

Unlike the previous sequence, the WRIVA `t05_v09_s00_r06_ArtifactInjection_M04_car` dataset evaluates the complementary failure mode of transient authentication. The scene contains a damaged vehicle that is a permanent component of the environment rather than a transient object. Therefore, a reliable transient authentication framework should preserve the vehicle and avoid producing any transient masks. Figure 7 compares *TRACE* with T-3DGS on this challenging sequence. Owing to the proposed SfM-guided authentication procedure, *TRACE* correctly determines that the scene contains no transient objects and consequently produces no transient masks. The resulting Gaussian Splatting reconstruction faithfully preserves the underlying scene geometry. In contrast, T-3DGS incorrectly identifies portions of the static scene as transient despite the absence of removable objects. These erroneous masks propagate directly into the optimization process, introducing unnecessary edits and visible blur in the rendered reconstruction. Although the damaged vehicle is correctly preserved, the false-positive masks degrade the quality of the surrounding static scene.

Together with the results in Fig. 6, these experiments demonstrate the importance of explicitly authenticating candidate objects before scene editing. Under challenging real-world capture conditions characterized by sparse viewpoint overlap and highly irregular camera trajectories, geometry alone can produce both false negatives (missed transient objects) and false positives (incorrectly removing persistent scene content). By combining semantic candidate recognition with multi-view

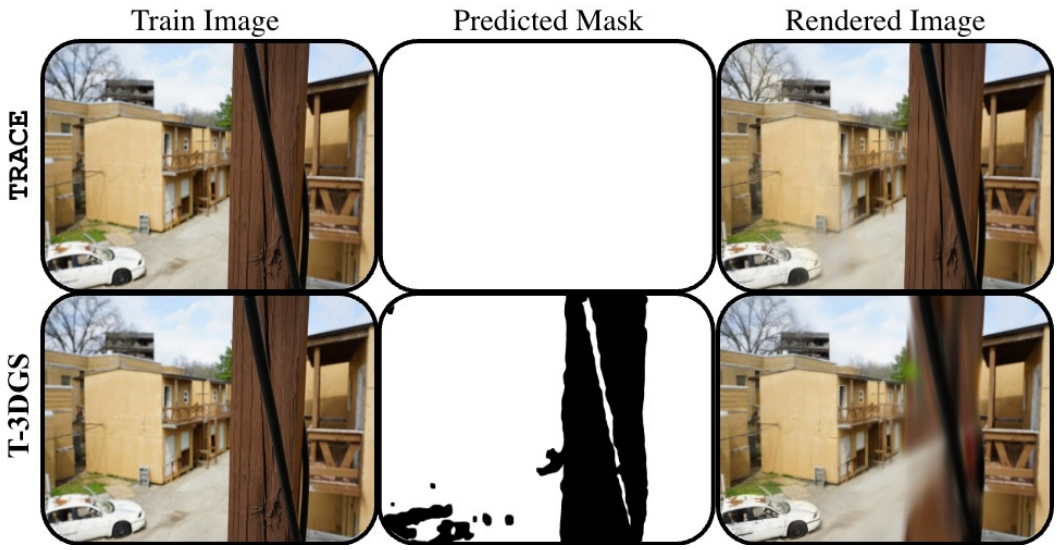


Figure 7: **False-positive transient detection on the WRIVA `t05_v09_s00_r06_ArtifactInjection_M04_car` sequence.** The damaged vehicle forms part of the permanent scene and should therefore be preserved. *TRACE* correctly authenticates the scene as containing no transient objects and consequently produces no transient mask, preserving the reconstructed geometry. In contrast, T-3DGS incorrectly classifies portions of the static scene as transient, resulting in erroneous masks and noticeable blur in the rendered reconstruction.

SfM-guided authentication, *TRACE* substantially improves the reliability of transient localization prior to Gaussian Splatting optimization.

We additionally observe that T-3DGS requires approximately 10 hours of optimization on a single NVIDIA RTX A5000 GPU for this sequence, yet still fails to reliably identify transient regions under these challenging real-world conditions. On the other hand, *TRACE* can be easily integrated into any GS based pipeline, and it would only take seconds to correctly obtain the masks. These results demonstrate that robust transient removal fundamentally depends on accurate transient authentication rather than geometric inconsistency alone.

A.2.2 Coverage-aware View Selection

This section presents the complete implementation details of the proposed coverage-aware view selection module within *TRACE*. While the main paper introduces the motivation and formulation of the proposed strategy, Algorithm 1 provides the exact implementation used throughout all experiments. The view selection module operates after the transient authentication stage. Given the authenticated transient masks $\mathcal{M}_{\text{tr}}$ (or any set of masks), the calibrated SfM reconstruction, and a user-defined view budget K, its objective is to identify a compact subset of training views $\mathcal{C}^* \subseteq \mathcal{C}$ with $|\mathcal{C}^*| \leq K$ that provides the strongest supervision for subsequent diffusion-guided inpainting and Gaussian Splatting optimization. Specifically, the algorithm takes as input

- the registered SfM cameras $\mathcal{C}$, together with their intrinsic and extrinsic parameters;
- the reconstructed sparse point cloud $\mathcal{P}$;
- the authenticated transient masks $\mathcal{M}_{\text{tr}}$;
- and the maximum number of editable views K.

The output is a compact set of informative cameras $\mathcal{C}^*$, which are subsequently selected for diffusion-guided inpainting before optimizing the Gaussian Splatting representation. The proposed mechanism is designed to satisfy two complementary objectives:

1. maximize multi-view geometric coverage of the reconstructed scene; and

Algorithm 1 Coverage Aware View Selection of *TRACE*

procedure SELECT*TRACE*(imgs.bin, pts3D.bin, mask_dir, K, α)
  Step 1: Parse SfM Data
  imgs ← READIMGS(imgs.bin)
  pts3D ← READPTS3D(pts3D.bin)
  Step 2: Build Coverage Maps
  cam_cov ← CAMCOVERAGE(imgs, pts3D) ▷ camera ↦ visible 3D points
  obj_cov, mask_area ← OBJECTCOVERAGE(imgs, pts3D, mask_dir) ▷ masked 3D points and mask pixel area
  Step 3: Initialize
  sel_cams ← [ ]
  cov_pts ← {} ▷ 3D points already covered
  Step 4: Greedy Selection Loop
  for $i \leftarrow 1$ to K **do**
    best_cam ← None
    best_score ← $-\infty$
    for all $cam \in$ cam_cov **do**
      if $cam \in$ sel_cams **then**
        continue
      end if
      new3D ← SIZE(cam_cov[cam] − cov_pts)
      maskScore ← mask_area[cam]
      score ← $(1-\alpha) \cdot$ new3D $+ \alpha \cdot$ maskScore
      if score $>$ best_score **then**
        best_score ← score
        best_cam ← cam
      end if
    end for
    if best_cam = None **then**
      break ▷ No more coverage gain
    end if
    APPEND(sel_cams)best_cam
    cov_pts ← cov_pts ∪ cam_cov[best_cam]
  end for
  return sel_cams
end procedure

2. prioritize views in which authenticated transient objects are most visible.

For every registered camera $c \in \mathcal{C}$, *TRACE* computes two complementary quantities. The first, $\mathrm{new}_{3D}(c)$, measures the number of previously uncovered SfM points contributed by camera c, encouraging complementary geometric coverage across the selected views. The second, $A(c)$, denotes the area of the authenticated transient mask in the image, favoring views that provide strong supervision for diffusion-guided inpainting. These two quantities are combined into the camera utility score

$$\mathrm{score}(c) = (1-\alpha)\,\mathrm{new}_{3D}(c) + \alpha\,A(c),$$

where $\alpha \in [0,1]$ balances geometric coverage and transient visibility. Starting from an empty camera set, *TRACE* greedily selects the camera with the highest score, updates the set of covered 3D points, and repeats the procedure until either K views have been selected or no additional geometric coverage can be obtained. Increasing K provides additional geometric supervision at the expense of editing more training images. Unless otherwise stated, Unless otherwise stated, we use $\alpha = 0.5$ throughout all experiments, assigning equal importance to geometric coverage and transient visibility during view selection. Algorithm 1 summarizes the complete coverage-aware view selection procedure, while the overall *TRACE* framework is shown in Figure 1.

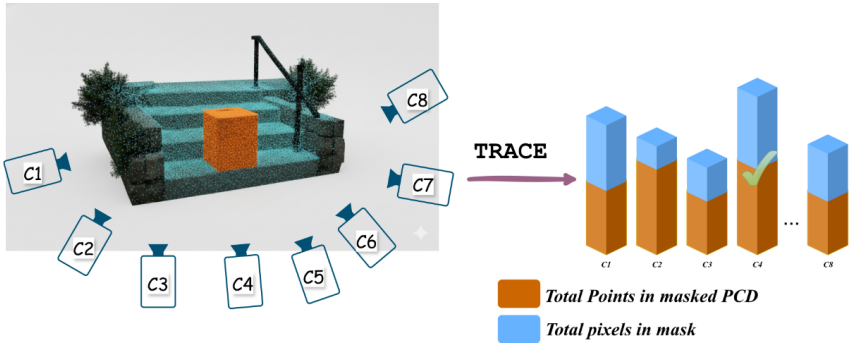


Figure 8: **Overview of the proposed coverage-aware view selection strategy in TRACE.** The proposed view selection module identifies a compact set of informative views for diffusion-guided inpainting. The left illustrates multiple camera views $(C_1, \ldots, C_8)$ observing an object (orange cube), where each view provides different geometric support in the underlying SfM reconstruction and different visibility in image space. For every camera, TRACE computes two complementary measures: (i) the number of newly covered 3D points contributed by the camera and (ii) the area of the authenticated transient mask. As illustrated on the right, cameras are ranked by jointly maximizing geometric coverage and visibility, allowing TRACE to select a minimal yet highly informative subset of views (e.g., C_4) for diffusion-guided inpainting. This geometry-aware selection provides stronger supervision for subsequent Gaussian Splatting optimization than conventional random view selection while requiring only a small number of edited images.

A.3 Ablation Studies

Here we provide ablations on the effect of K on the results for Neural 3D, and WRIVA datasets.

A.3.1 On Neural 3D dataset

Figure 9 illustrates how the number of inpainted views K influences the quality and consistency of object removal in the *coffee martini* sequence. Here, the target object is the painting mounted on the cupboard, which we aim to remove using *TRACE* applying to InFusion. We compare two configurations, $K = 2$ and $K = 3$, and visualize the resulting renderings from three different viewpoints. When only two views are inpainted ($K = 2$; top row), several viewpoints still exhibit noticeable rendering artifacts around the removed region, such as blurred geometry, inconsistent shading, or incomplete background reconstruction. These artifacts arise because two inpainted observations are not sufficient to fully constrain the appearance of the newly exposed surfaces across all viewing directions.

On the other hand, when the number of inpainted views is increased to three ($K = 3$; bottom row), the improvement is substantial. Nearly all viewpoints show coherent reconstruction of the previously occluded region, with cleaner textures, improved geometric consistency, and significantly fewer artifacts. This demonstrates that increasing K provides more reliable multi-view consistency for the inpainted region, enabling the model to better propagate plausible content throughout the 3D Gaussian representation.

Overall, this example highlights the importance of selecting an adequate number of inpainted views: too few lead to view-dependent artifacts, whereas a slightly larger K ensures more stable and photorealistic object removal across the scene.

A.3.2 On WRIVA dataset

Following the procedure described in the main paper, we also evaluate the removal of another parked car (transient) located in front of the building. We first track the car across the entire sequence, and fig. 10 shows all images in which this object appears. Applying *TRACE* to this set reveals how the algorithm identifies the most informative views for inpainting.



Figure 9: **Effect of K on the rendered results:** As shown in the top row, when selecting two cameras results in artifacts within the rendered inpainted region. However, for the *cofee_martini* dataset, $K = 3$ demonstrates improved performance across nearly all viewpoints

Setting	**DreamSim** $\downarrow$	**PSNR (dB)** $\uparrow$
Random with Object	0.1717	17.408
1 View	0.1671	17.605
2 Views	0.1633	17.494

Table 4: Comparison of DreamSim and PSNR across different view-selection strategies.

When $K = 1$, the method selects the second image in the first row of fig. 10. This image appears to be the image that obtains the highst score, which implies the balance between total 3d points the masked area has and the total mask pixels. When $K = 2$, *TRACE* selects the two images in the second column; these views complement each other in terms of geometric coverage, capturing different subsets of the object's visible surfaces. When $K = 3$, the algorithm adds the last image in the third row, which provides an additional viewpoint that captures regions of the car that are not visible in the first two selections. It is important to emphasize that the selection made by *TRACE* is driven not merely by mask size, but also by the number of registered 3D points associated with the object in each view. As a result, the method naturally identifies a set of *disjoint, basis-like* viewpoints that span the object's appearance and geometry. These views collectively cover the car from different angles, providing sufficient constraints for robust inpainting and consistent scene reconstruction after object removal. This kind of structured selection would not arise from a random choice of inpainting views. Random sampling may easily pick redundant views with similar viewpoints, producing insufficient coverage of object geometry and leading to noticeable artifacts in the final reconstruction. In contrast, *TRACE* ensures that each selected camera contributes new and complementary geometric information, making the inpainting process more effective and the final rendered scene significantly more stable and coherent. The test rendered results are shown in Figure 11, and the average metrics for the entire test dataset is shown in Table 4.

A.4 Additional Qualitative Results

A.4.1 Spin-NeRF

Figure 12 presents additional qualitative comparisons on the Spin-NeRF dataset. GScream removes the object but often introduces blurred surfaces and smeared textures, particularly around depth-discontinuous regions such as the staircase edges and the tree trunk. InFusion seems to be producing sharper outputs than GScream; however, it frequently suffers from structural distortions. For



Figure 10: **Tracked Object Across the Scene.** The figure illustrates the trajectory of the tracked object. This setting poses a significant challenge for determining which specific instance should be inpainted, as the object appears from irregular viewpoints and is not fully visible in many of the frames. In such cases, our proposed mask-score metric becomes especially valuable, enabling reliable identification of the appropriate images for inpainting even when object visibility is inconsistent.



Figure 11: **Effect of K on Rendered Results.** The figure compares GScream with our proposed *TRACE* augmentation for different values of K. Instead of selecting a random image containing the transient object for inpainting, *TRACE* optimally selects views that maximize object coverage and geometric influence. As shown, this targeted selection significantly reduces residual artifacts in the test renderings. A closer inspection of the bottom-right region of each view, where the removed vehicle originally appeared, reveals that GScream alone leaves subtle inconsistencies, while *TRACE* increasingly suppresses these artifacts as K increases. Notably, *TRACE* with even $K = 1$ provides a substantial improvement, and $K = 2$ produces results visually closest to the ground truth, demonstrating the effectiveness of our informed view-selection mechanism.

instance, color bleeding on the stairs, warped geometry behind the orange-colored net, and flattened or inconsistent textures along the grass and bark. In contrast, our method (on top of GScream) preserves geometric structure and high-frequency details far more reliably. On the staircase region, our approach reconstructs clean and sharp step boundaries without the ghosting or color drift seen in the baselines. Around the orange-colored net region, our method maintains coherent background content without the severe warping present in InFusion. Finally, around the tree trunk, our reconstruction avoids the smudging from GScream and the patchy artifacts from InFusion, producing a natural and consistent appearance closely matching the ground truth. These results highlight that our *TRACE*–based pipeline generalizes well to diverse scenes and consistently achieves higher visual fidelity across the Spin-NeRF dataset.

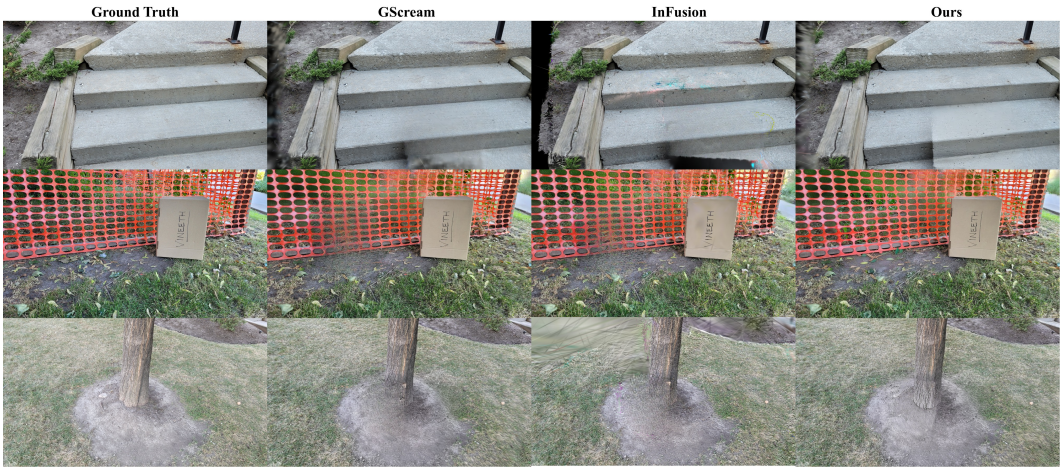


Figure 12: **Qualitative comparison on the Spin-NeRF dataset.** As can be seen, GScream removes the object but often produces blurred surfaces and smeared textures, particularly on the staircase and near the tree trunk. InFusion improves sharpness but frequently introduces structural distortions, such as color bleeding on the stairs, warped geometry behind the orange fence, and inconsistent textures around the grass and bark. In contrast, our method preserves geometric structure and high-frequency details far more consistently. Across all regions, our reconstruction remains closest to the ground truth, avoiding the ghosting, warping, and texture flattening found in competing methods. These results demonstrate the robustness and superior visual fidelity of our approach on the Spin-NeRF dataset.

A.4.2 NEURAL 3D DATASET

Figure 13 presents qualitative results on the Neural 3D dataset, where the objective is to remove the puppy sitting on the chair across two different timestamps and camera viewpoints. As can be seen InFusion is able to suppress the visible puppy, but it introduces severe distortions—such as warped chair geometry, smeared textures, and ghost-like remnants of the original object (highlighted in red). These artifacts vary significantly across timestamps and viewpoints, indicating a lack of temporal and multi-view consistency. On the other hand, our method achieves clean object removal while faithfully reconstructing the underlying background. The chair surface, shadows, and surrounding objects remain structurally coherent and visually stable across both timestamps. The background is reconstructed without the tearing, blurring, or color drift observed in InFusion. Importantly, our results remain consistent across camera views, demonstrating strong cross-view coherence and robustness to motion in the scene. Overall, our approach produces the most natural and artifact-free reconstruction for this challenging dynamic object-removal task.

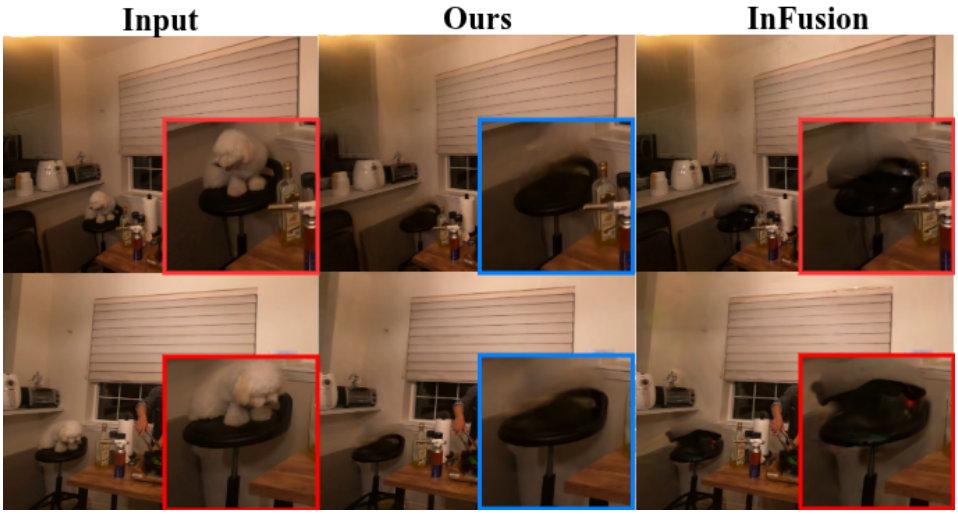


Figure 13: **Object removal results on the Neural 3D dataset.** The goal in this scene is to remove the puppy sitting on the chair across two different timestamps and camera viewpoints. While InFusion succeeds in suppressing the puppy, it introduces severe distortions, including warped chair geometry, smeared textures, and ghost-like residuals of the object, which vary noticeably across views and time. In contrast, our method produces clean and stable background reconstructions, preserving the chair structure, shading, and surrounding objects without introducing blurring or tearing artifacts. Importantly, our results remain consistent across both timestamps and viewpoints, demonstrating strong temporal and cross-view coherence. Overall, our approach delivers the most natural and artifact-free reconstruction for this challenging dynamic-object removal task.