

TRACE: Instance-Level Open-Vocabulary Inventory Generation for 3D Forensic Evidence Reconstruction

Florian Eichinger^{1,2}, Michael Greza^{1,2}, Benjamin Busam^{1,2}

¹ Photogrammetry and Remote Sensing, TU Munich, Germany - (florian.eichinger, michael.greza, b.busam)@tum.de

² Munich Center for Machine Learning, Germany - (florian.eichinger, michael.greza, b.busam)@tum.de

Keywords: 3D Reconstruction, Multiview Instancing, Spatial Inventorization, Language-Aligned Semantics.

Abstract

State of the art crime scene documentation not only relies on visually and geometrically faithful 3D reconstruction, but also structured access to potentially relevant evidentiary objects. Existing language-aware 3D scene representations enable open-vocabulary object querying and scene-level reasoning, yet they do not explicitly tackle the formation of globally consistent physical object instances across views. This limitation is particularly critical in forensic settings, where evidence is tied to discrete objects and where scenes are often visually complex, unstructured and atypical in appearance. We present *TRACE*, a training-free framework for instance-level open-vocabulary inventory generation in 3D forensic evidence reconstruction. *TRACE* combines multiview instance formation, language-aligned aggregation, and 2D–3D semantic lifting within a unified Gaussian-based scene representation. It therefore moves beyond semantic scene querying towards a structured and language-addressable 3D inventory of physical evidence. The proposed formulation provides a bridge between investigator driven object assessment and instance-aware 3D scene understanding for forensic analysis.

1. INTRODUCTION

The documentation of crime scenes is inherently a spatial and evidentiary task. Investigators must not only record which potentially crime-relevant objects are present, but also preserve their spatial arrangement and contextual relationships. In practice, this process still relies heavily on photographs, sketches, notes and inventory-like records created during the scene inspection. At the same time, forensic literature has emphasized the value of digital 3D reconstruction for non-destructive and reviewable documentation of crime scenes, while also framing the broader integration of 3D imaging and modelling as an increasingly important direction in forensic science (Carew et al., 2021; Balbudhe et al., 2025).



Figure 1. *TRACE*: From prompt-based multiview observations to a structured, instance-level 3D inventory of physical evidence.

Despite these advances, there remains a gap between the geometric scene capture and semantic evidence organization. Modern 3D reconstruction approaches such as NeRF (Mildenhall et al., 2021) and 3D Gaussian Splatting (Kerbl et al., 2023) can produce detailed scene representations in terms of photorealism and geometric fidelity. However they do not directly yield a spatially grounded and interactive inventory of its contained physical objects. For practical forensic analysis, a useful digital scene representation requires to move beyond pure reconstruction towards semantically structured and object-centric scene models.

Recent advances in the field of 3D vision-language integration provide an important foundation for addressing this challenge. Language grounding in 3D scenes has first been demonstrated for radiance field based representations (Kerr et al., 2023) and was subsequently transferred to 3D Gaussian Splats (Qin et al., 2024), enabling open-vocabulary querying directly in 3D. At the same time, recent approaches have begun to explore generalizable models for 3D Gaussian Splat (3DGS) scene understanding that operate natively on Gaussian primitives and learn semantic scene representations from large-scale 3DGS datasets (Li et al., 2025). Together, these developments make 3DGS an attractive spatial anchor not only for visual reconstruction but also for semantic augmentation and interactive scene understanding. However, for forensic applications, semantic scene querying alone is not sufficient. Evidence is handled at the level of concrete objects rather than semantic categories. In practical scenarios, the representations require the ability to resolve distinct objects and preserve their identity across views. This is particularly relevant as forensic environments are complex, unstructured and atypical in appearance. Potentially relevant evidence may therefore stem from rare object categories with unusual object states and uncommon spatial constellations. In such settings, purely scene-centric approaches based on broad learned scene priors may be less well matched to the demands of forensic analysis, especially when evidentiary relevance is driven by atypical object-level observations rather than domi-

nant scene semantics (Li et al., 2025). This motivates a training-free and instance-centric formulation within which object identity is established directly from the observed scene geometry and visual appearance.

To this end, we propose *TRACE*, a training-free framework for instance-level inventory generation in 3D. Starting from high-resolution RGB imagery, *TRACE* transforms prompt-based 2D object observations into a semantically searchable and spatially grounded 3D inventory. The central stage is a multiview-consistent instancing procedure that combines scene geometry with visual appearance to assign globally stable identifiers to distinct same-class objects across views. These instance-level observations are then aggregated into language-aligned descriptors and lifted into a 3DGS. This yields an enriched scene that jointly encodes geometry, semantic accessibility and object identity.

2. RELATED WORK

Recent years have seen rapid progress in enriching neural scene representations with semantics and language-aligned descriptors, posing various possibilities also for forensic applications.

Language Grounding in 3DGS. A first line of work investigates the transfer of semantics or language information extracted in 2D into Gaussian-based 3D scene representations. *LangSplat* (Qin et al., 2024) introduces language-aligned Gaussian fields for efficient open-vocabulary querying in 3D scenes by learning a language field directly in Gaussian space. In a related learning-based schema, *3D Vision-Language Gaussian Splatting* (Peng et al., 2024) improves semantic consistency through cross-modal rasterization and view blending, further demonstrating the potential of semantically promptable Gaussian scene representations. More recent works such as Occam’s LGS (Cheng et al., 2024) and Dr.Splat (Jun-Seong et al., 2025) have also explored training-free alternatives for semantic transfer. They improve both runtime efficiency and per-point semantic consistency but lack occlusion handling – especially posing a problem in unstructured and chaotic environments such as crime scenes. VALA (Wang et al., 2025) addresses these issues by introducing the concepts of visibility-aware gating and a streaming cosine-median feature aggregation to ensure that only truly visible Gaussians contribute to the semantic representation. *In contrast, TRACE moves beyond semantic transfer alone by explicitly forming globally consistent object instances before lifting language-aligned descriptors into 3D.*

Instance-Level Reasoning in 3DGS. A second line of work moves closer to object-level reasoning and segmentation in Gaussian scenes. Rather than focusing primarily on feature-aligned scene semantics, these approaches aim to identify and group object-level regions directly in 3DGS. *Gaussian Grouping* (Ye et al., 2024) thereby introduces the segment-anything paradigm in 3D Gaussian Splats by lifting SAM guided 2D information in 3D for segmentation, grouping and editing via identity encodings. *SAGA* (Cen et al., 2025) augments Gaussians with learned affinity features distilled from 2D SAM masks to enable promptable segmentation in 3D. More recently, *OpenSplat3D* (Piekenbrinck et al., 2025) combines instance-aware scene partitioning with language embeddings for open-vocabulary 3D instance segmentation. *These methods are particularly relevant to TRACE as they move beyond scene-level semantics toward object-aware reasoning in 3DGS. TRACE, however, is not primarily designed for promptable 3D segmentation*

or mask generation, but for constructing a globally consistent 3D instance-inventory with language addressable entries following a training-free paradigm.

Scene-Centric 3D Vision-Language Models. Complementary to the above works, recent approaches have begun to formulate global scene understanding directly on top of Gaussian primitives through learned 3DGS representations. *GaussianVLM* (Halacheva et al., 2025) introduces a scene-centric 3D VLM that directly associates language with Gaussians for holistic text-based scene interaction. *While being particularly promising for general 3D scene understanding, broad scene-level priors may be less directly aligned with forensic inventory generation: Crime scenes tend to be complex, unstructured and atypical in both appearance and spatial arrangement. In such settings, broad scene-level priors may be less informative than explicit per-scene and object-centric formulations.*

Building on these developments, *TRACE* combines training-free instance formation with language-aligned 3D scene enrichment. Unlike existing approaches, which primarily focus on semantic transfer, promptable segmentation or scene-level reasoning, *TRACE* first establishes globally consistent object instances and only then lifts language-aligned descriptors in 3D to generate structured forensic inventories.

3. METHODOLOGY

The proposed framework aims to generate a semantically searchable, instance-level inventory from high-resolution RGB-imagery of crime scenes. Starting from prompt-based 2D object hypotheses, the methodology combines multiview-consistent instancing with language-driven segmentation and inventory generation in 3D Gaussian Splats. Figure 2 illustrates the overall workflow of the proposed method.

3.1 Multiview-Consistent Instancing

Given a set of binary 2D object masks $\mathcal{M} = \{m_i^k\}$ extracted independently from the input views I_i , our goal is to assign a globally consistent instance identity to masks of the same semantic class. We formulate this as a graph-based association problem requiring the consideration of both scene geometry and visual object appearance on top of a photogrammetric reconstruction. To this end we employ COLMAP (Schönberger and Frahm, 2016; Schönberger et al., 2016) to obtain calibrated cameras, a sparse SfM model and dense multiview depth maps. In addition, a foundational self-supervised encoder, DINOv2 (Oquab et al., 2023) resp. DINOv3 (Siméoni et al., 2025), is used to extract object-level appearance descriptors from masked object regions.

As an initial preprocessing step, prompt-based 2D object masks are generated independently in each input view using SAM3 (Carion et al., 2025). Given a class specific concept prompt such as "knife", the model produces one binary mask per detected object instance together with an associated confidence score. *TRACE* employs SAM3 solely as a promptable segmentation front-end to obtain class-specific 2D mask hypotheses, which are subsequently processed by the proposed multiview association framework.

For each mask m_i^k all valid pixels with dense depth are back-projected into 3D, yielding a mask anchor $A_i^k = \{P_i^k, c_i^k, B_i^k\}$, where $P_i^k \subset \mathbb{R}^3$ is the 3D support point set, c_i^k its centroid and

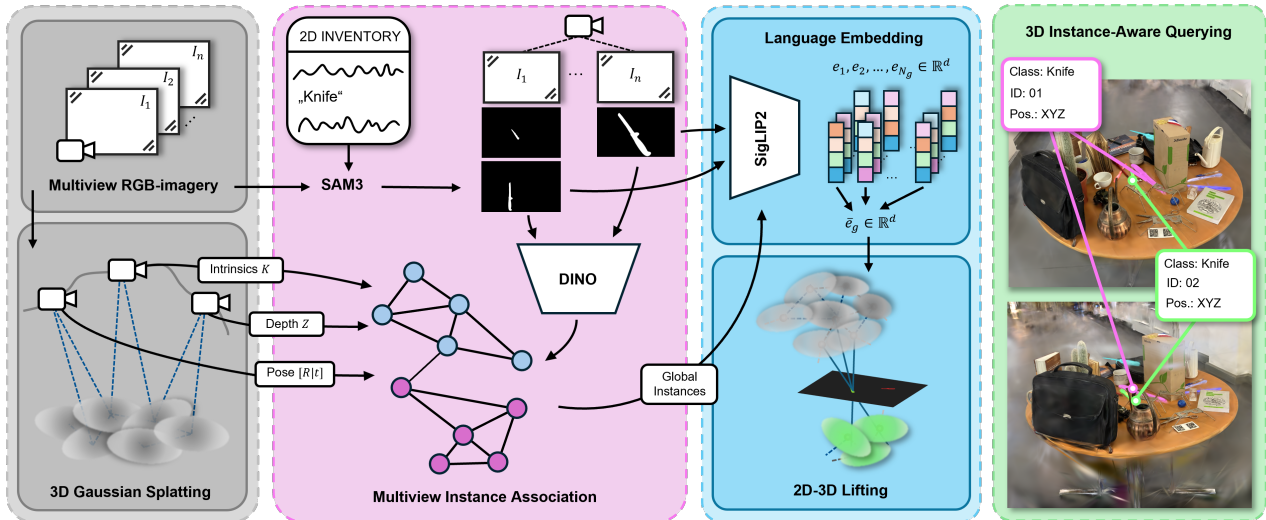


Figure 2. *TRACE* pipeline overview. Starting from multiview RGB imagery, prompt-based 2D object masks are extracted with SAM3 and associated across views via geometry-aware and appearance-aware multiview instancing. Based on COLMAP geometry and DINOv2/v3 descriptors, the proposed framework establishes globally consistent same-class object identities across the scene. The resulting global instances are then encoded with SigLIP2 to obtain language-aligned instance descriptors and subsequently lifted into a 3D Gaussian Splat representation by assigning instance-level semantics to geometrically supported Gaussian subsets. This yields an enriched 3D scene representation that jointly preserves spatial structure, object-level identity, and language-accessible semantics, thereby enabling instance-aware open-vocabulary querying in 3D.

B_i^k an axis-aligned 3D bounding box. This procedure converts each 2D hypothesis into a geometrical observation of an underlying physical object. To avoid quadratic matching over all images we construct a lightweight image-neighborhood graph: Two images are only considered for the matching if they are neighbors in capture order (± 4 frames) or among the k nearest camera centers in 3D ($k = 4$). Within an admissible image pair (I_i, I_j) , two anchors A_i^k and A_j^ℓ are retained as candidate pairs if they are either geometrically admissible according to Eq. (1) or linked by a top-1 appearance relation across the image pair. Geometrically admissible pairs are identified by

$$\|c_i^k - c_j^\ell\|_2 \leq \alpha \max\{\text{diag}(B_i^k), \text{diag}(B_j^\ell)\}, \quad (1)$$

with $\text{diag}(B)$ denoting the euclidean diagonal distance of the 3D bounding box. This candidate gating strategy removes geometrically implausible pairs early while preserving long-range support through the neighborhood graph.

In parallel, each instance is additionally assigned an appearance descriptor $f_i^k \in \mathbb{R}^d$ by encoding the masked image crop with DINOv2/v3. The appearance based similarity between two candidate anchors (A_i^k, A_j^ℓ) is then defined as

$$s_{ij}^{\text{app}} = \frac{1 + \cos(f_i^k, f_j^\ell)}{2}, \quad (2)$$

which maps cosine similarity to the interval $[0, 1]$. To remove ambiguous matches between visually similar objects of the same semantic class, appearance is not only evaluated by its absolute similarity but also by its discriminative strength within the target image through the margin between the best and the second-best appearance candidate.

For each surviving directed candidate pair $A_i^k \rightarrow A_j^\ell$ the source support P_i^k is projected into the target view I_j . In contrast to a purely projection-count based formulation, *TRACE* evaluates geometric consistency in an occlusion-aware manner. For every

projected source point, the projected depth is compared against the target-view depth map. A projected point is considered potentially visible if it is not significantly further away than the observed target depth at the corresponding pixel. Points that fall behind the target surface are treated as occluded and therefore do not negatively contribute to the geometric match. This is particularly important in cluttered crime scenes, where the same physical object may be fully visible in one view but heavily occluded in another.

Based on this occlusion-aware projection, we compute three directed geometric (sub-)scores:

1. s_{hit} , measuring the fraction of projected source points that are inside the target mask relative to the subset of source points that are potentially visible in the target view;
2. s_{cov} , measuring how much of the target mask is covered by the visibility-filtered projected source support;
3. s_{depth} , measuring the consistency between projected source depth and observed target depth.

The directed geometric score is then computed through the weighted sum

$$s_{i \rightarrow j}^{\text{geom}} = \lambda_1 s_{i \rightarrow j}^{\text{hit}} + \lambda_2 s_{i \rightarrow j}^{\text{cov}} + \lambda_3 s_{i \rightarrow j}^{\text{depth}}, \quad (3)$$

with $(\lambda_1, \lambda_2, \lambda_3) = (0.45, 0.35, 0.20)$ in all experiments. Mutual consistency is established by the symmetrization of the directed graph to consider also the reverse projection of P_j^ℓ into the original source view I_i yielding $s_{j \rightarrow i}^{\text{geom}}$.

Rather than using geometry and appearance in a sequential manner, both cues are already fused at the directed edge level. Let $r_{i \rightarrow j}^{\text{geom}}$ therefore denote a geometric reliability term derived from the amount of valid projected support, and let $r_{i \rightarrow j}^{\text{app}}$ denote

an appearance reliability term derived from the appearance margin. The fused directed score is then given by

$$s_{i \rightarrow j} = \frac{r_{i \rightarrow j}^{\text{geom}} s_{i \rightarrow j}^{\text{geom}} + \beta r_{i \rightarrow j}^{\text{app}} s_{i \rightarrow j}^{\text{app}}}{r_{i \rightarrow j}^{\text{geom}} + \beta r_{i \rightarrow j}^{\text{app}} + \varepsilon}, \quad (4)$$

where β denotes the relative contribution of the appearance. In this formulation, geometry remains the major cue, while appearance becomes particularly influential in cases of incomplete depth support or spatial proximity of the instances. The final symmetric edge weight is computed as

$$\omega_{ij} = \text{HM}(s_{i \rightarrow j}, s_{j \rightarrow i}), \quad (5)$$

where $\text{HM}(\cdot)$ denotes the harmonic mean which penalizes asymmetry or partially inconsistent matches.

Global instance identities $\mathcal{G} = \{g_i\}$ are finally obtained by constrained graph clustering using Union Find: Two connected components are merged only if they do not already contain anchors from the same image as one object can physically only exist once per view. This ultimately yields globally consistent same-class object instances across the entire scene by combining physically grounded geometry with appearance-based disambiguation.

3.2 Language-Guided Instance Encoding

For a given global instance g , all contributing mask observations from the multiview association graph are considered jointly. For each mask m_i^k , the corresponding image region is cropped to the object while the background is entirely suppressed. These masked object crops are then encoded independently using SigLIP2 (Tschannen et al., 2025) image encoder to obtain per-mask normalized feature vectors. In contrast to the previous instancing stage, where appearance is used to disambiguate geometric correspondences, the objective here is to derive a single robust descriptor for each already established global instance.

Let $\mathcal{V}_g = \{v_1, \dots, v_{N_g}\}$ denote the set of mask observations associated with the global instance g and let $e_n \in \mathbb{R}^d$ be the corresponding SigLIP2 feature vector of observations v_n . Since not all observations are equally reliable, we employ the SAM3 per-mask model confidence to serve as reliability score w_n . The global instance embedding is then computed as the weighted average

$$\tilde{e}_g = \frac{\sum_{n=1}^{N_g} w_n e_n}{\sum_{n=1}^{N_g} w_n}, \quad \bar{e}_g = \frac{\tilde{e}_g}{\|\tilde{e}_g\|_2}. \quad (6)$$

This yields a single robust, language-aligned feature vector $\bar{e}_g \in \mathbb{R}^d$ for each global instance.

3.3 2D–3D Feature Lifting

To transfer these instance-level features into 3D, the 3D Gaussian Splat is used as a spatial carrier structure. For each global instance, a subset of supporting Gaussians is identified by aggregating multiview instance evidence across the associated mask observations under geometric consistency constraints.

Let q denote a Gaussian primitive and let $v_n \in \mathcal{V}_g$ be one observation of instance g . For each pair (q, v_n) , the Gaussian center is projected into the corresponding image view. A contribution

is only considered if the projection is valid, lies inside the image domain, and falls within the corresponding instance mask. If these conditions are satisfied, a non-negative view-specific support score $s_n(q)$ is assigned, which depends on the consistency between projected Gaussian depth and observed image depth. This yields one support contribution per observation and Gaussian. The accumulated support of Gaussian q for instance g is then defined as

$$S_g(q) = \sum_{n=1}^{N_g} s_n(q), \quad N_g^{\text{sup}}(q) = \sum_{n=1}^{N_g} \mathbf{1}[s_n(q) > 0]. \quad (7)$$

Here $S_g(q)$ denotes the aggregated soft support score and $N_g^{\text{sup}}(q)$ is the number of supporting observations. The final lifted Gaussian subset for instance g is obtained by thresholding both quantities:

$$\mathcal{Q}_g = \{q \mid S_g(q) \geq \tau_S \wedge N_g^{\text{sup}}(q) \geq \tau_N\}, \quad (8)$$

with τ_S and τ_N denoting the minimum soft-support and minimum observation-support thresholds. The retained Gaussians are then assigned two additional attributes: a global instance identifier and the SigLIP2 feature vector $\bar{e}_g$ of the corresponding instance. In this way, each lifted Gaussian inherits both object membership and a language-aligned semantic descriptor. Finally, the lifted subsets of all global instances are merged into a single enriched 3D representation which augments the original 3DGS scene with instance-aware and semantic attributes and thereby enables subsequent open-vocabulary querying directly in 3D.

4. EXPERIMENTS

4.1 Experimental Setup

We quantitatively assess the proposed multiview instancing approach by evaluating it on three controlled small-scale tabletop datasets of increasing semantic and geometric complexity. All datasets are reconstructed from calibrated multiview imagery and annotated at the level of individual 2D mask observations. The experiments are designed to analyze not only the overall performance of the proposed method but also the relative contribution of geometric consistency and visual appearance.

We chose knives as the primary object class within the datasets both because they fit the forensic evidence setting and because they are visually challenging due to their highly reflective surface. Dataset 1 contains eight randomly distributed and partially overlapping distinct kitchen knives. Dataset 2 increases ambiguity by replacing half of the knives with visually identical ones, thereby testing the role of geometry in disambiguating repeated instances of one category. The third dataset introduces a second object class together with strong structured occlusions allowing us to assess the robustness of the method under reduced inter-view connectivity and class-dependent visibility. Datasets 1 and 2 are designed to isolate the relative roles of geometry and appearance under strong inter-view connectivity. Dataset 3 tests whether the proposed fusion remains beneficial when geometric continuity is disrupted by structured occlusions.

We compare the full combined method against geometry-only and appearance-only variants. For the appearance-based variants both DINOv2-B and DINOv3-B are considered. All ex-

periments are evaluated using the same clustering pipeline and identical hyperparameter settings unless stated otherwise.

In addition to the quantitative evaluation of multiview-consistent instancing, we also provide qualitative visualizations for each dataset after the downstream instance encoding and 2D–3D lifting stages. These visualizations illustrate how the recovered global instances can be transferred into the Gaussian-based scene representation and subsequently accessed through category-level 3D prompting. In this way, the experiments assess not only the clustering accuracy of the proposed association framework, but also its practical suitability as the basis for instance-aware and language-addressable 3D inventory generation.

4.2 Metrics

The predicted global instance clusters are compared against manually defined ground truth associations provided for each 2D mask observation. We formulate the evaluation as a clustering problem over all mask instances and report Adjusted Rand Index (ARI), B^3 Precision, Recall, and F1, as well as homogeneity, completeness, V-measure, and the difference between the predicted and true number of clusters. In addition, we report the ratio between predicted and ground-truth numbers of clusters ($\#Pred/\#GT$) in order to explicitly capture potential over-/under-segmentation effects.

These metrics provide complementary perspectives on clustering quality. ARI measures the overall agreement between predicted and ground-truth partitions while correcting for chance. B^3 Precision, Recall, and F1 quantify clustering quality at the level of individual instances and are therefore particularly suitable for evaluating multiview association consistency. Homogeneity, completeness, and V-measure provide an additional characterization of cluster purity and fragmentation. Full metric tables (tab. 3, 4, 5, 6) including B^3 Precision/Recall, homogeneity, completeness, and V-measure are provided in the Appendix.

5. RESULTS

Dataset 1 - varying knives. On Dataset 1 (Tab. 1), both the combined method and the geometry-only variant achieve perfect clustering performance. This result is consistent with the acquisition setup and should be understood as an almost ideal case for geometry-based association. The sequence consists of a full circular capture around the table, all eight knives remain visible throughout the trajectory, strong occlusions are absent, and SAM3 detects each knife as a separate instance in every view. As a consequence, the 3D support of each object remains well connected across the entire view graph such that geometric consistency alone is sufficient to recover the correct global instance structure. The perfect performance of the combined variant therefore does not indicate an additional gain from appearance in this setting, but rather confirms that geometry is already fully informative.

Appearance-only clustering remains weaker. However, the results do not reflect a general instability to separate different knives. Instead the dominant error is mild over-segmentation in the form of small fragments and singleton clusters. This is reflected by the increased number of predicted clusters (15/8 resp. 10/8), while the clustering scores remain high overall with ARI values of 0.987 and 0.927. This allows us to conclude that

the dominant error is not general confusion between different objects but incomplete reassociation of the same object across strongly varying viewpoints. A plausible explanation is that the circular acquisition, while highly favorable for geometry, is less favorable for pure appearance, since the same knife is observed under large pose changes and partially reversed orientations over the full trajectory.

Dataset 2 - equal knives. On Dataset 2 (Tab. 1), the general acquisition setup remains identical to Dataset 1, again yielding perfect clustering scores for both the combined method and the geometry-only variant. The crucial difference to Dataset 1 is therefore not geometric but semantic as we introduce four visually nearly identical knives.

This modification primarily affects the appearance-only variants. The degradation becomes substantially stronger on Dataset 2 with ARI dropping to 0.648 resp. 0.750. At the same time the number of predicted clusters increases (18/8 resp. 12/8), indicating pronounced over-segmentation. This behaviour is consistent with the intended stress test: By introducing visually interchangeable instances, appearance alone is no longer sufficient to reliably maintain instance identity across views. The DINOv3-B backbone mitigates this effect noticeably, but does not resolve it entirely.

The qualitative results for both Dataset 1 and Dataset 2, represented in Fig. 3, are consistent with the quantitative findings.

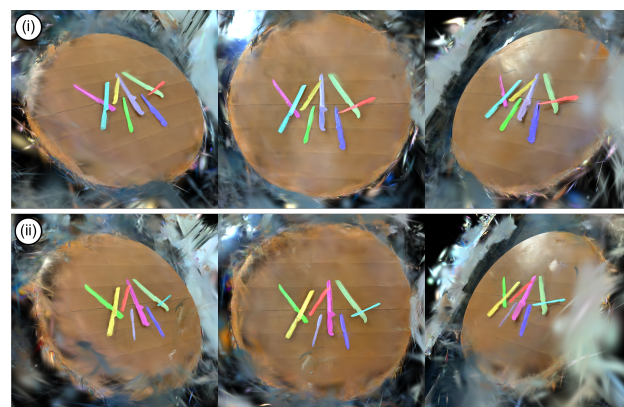


Figure 3. Qualitative results for the single-class knife datasets. (i): Dataset 1 & (ii) Dataset 2. Colors indicate recovered global instance identities after multiview association for the prompt "Knife".

For both datasets, the lifted Gaussians form spatially coherent and instance-separated 3D object hypotheses for the queried category "Knife". In Dataset 1, this reflects the almost ideal conditions already observed in the clustering analysis, where geometric continuity across views is sufficient to preserve all eight instances. In Dataset 2, the visualization remains qualitatively well structured despite the presence of visually nearly identical knives, supporting the conclusion that the proposed geometry-aware association stabilizes instance formation even when appearance alone becomes ambiguous. The figure therefore illustrates that global identities are not only quantitatively valid at the clustering level, but also sufficiently consistent to support instance-aware 3D prompting in the reconstructed scene.

Dataset 3 - occluded multi-class. On Dataset 3 (Tab. 2), the role of geometry and appearance changes significantly in comparison to Datasets 1 and 2. While the previous datasets were

Table 1. Quantitative evaluation on Datasets 1 and 2. Dataset 1 contains eight distinct knives, while Dataset 2 replaces half of the objects with visually near-identical knives. Geometry-driven variants remain perfect on both datasets, whereas appearance-only clustering degrades notably on Dataset 2.

Method	Dataset 1: varying knives				Dataset 2: identical knives			
	ARI	B ³ F1	MVA	#Pred/#GT	ARI	B ³ F1	MVA	#Pred/#GT
Combined	1.000	1.000	1.000	8/8	1.000	1.000	1.000	8/8
Geometry-only	1.000	1.000	1.000	8/8	1.000	1.000	1.000	8/8
Appearance-only (DINOv2-B)	0.927	0.927	0.978	15/8	0.648	0.675	0.813	18/8
Appearance-only (DINOv3-B)	0.987	0.987	0.996	10/8	0.750	0.777	0.856	12/8

recorded under the same acquisition setup without strong discontinuities, dataset 3 introduces a cross-shaped partition splitting the table into four mutually occluded sectors. Due to reduced inter-view connectivity, geometry alone is no longer capable to recover the correct global instance structure. This is directly reflected by the drop in ARI to 0.904 (globally) as well as the over-segmentation in 14/9 clusters. The resulting error remains purely fragmentary as indicated by the perfect precision and homogeneity scores while the lower recall shows that some object instances can no longer be kept connected across the full sequence.

At the same time, Dataset 3 demonstrates that the relative contribution of geometry and appearance changes once the scene is no longer fully connected across views. For mugs, the combined method achieves perfect clustering, whereas geometry-only remains slightly fragmented and both appearance-only variants are clearly weaker. For knives, however, appearance-only with DINOv3-B reaches perfect clustering, while geometry-only and the combined variant remain below this optimum. A plausible explanation is that the setting of cross-shaped partitions and self-occluding objects introduce different occlusion patterns across object instances, thereby interrupting the geometric continuity on which geometry-only association relies. In such cases, appearance becomes essential for reconnecting object identity across views. Importantly, the best global result is still obtained by the combined method, showing that appearance is not merely an auxiliary signal but a necessary complement once purely geometric association becomes fragmented. Dataset 3 shows the necessity to include appearance: unlike Datasets 1 and 2, it demonstrates that geometry-only is not universally sufficient under structured occlusion, and that appearance can substantially stabilize the global instancing result when geometric support becomes incomplete.

The qualitative results shown in Fig. 4 highlight another important property of the proposed downstream lifting stage. Although the clustering results on Dataset 3 are not perfect, the final 3D instance visualization is qualitatively correct for both object categories. This behaviour can be explained by the fact that the dominant residual error is fragmentary over-segmentation rather than systematic cross-instance confusion. In cases where a single object is represented by two spatially overlapping clusters, the downstream lifting and scene prompting procedure follows a winner-takes-most paradigm: the dominant cluster accumulates consistent multiview support across the relevant observations, whereas small fragments receive insufficient support under spatial overlap and are therefore not kept in the final lifted representation. As a consequence, minor cluster splits that remain visible in the instancing metrics do not necessarily propagate into the final 3D inventory.

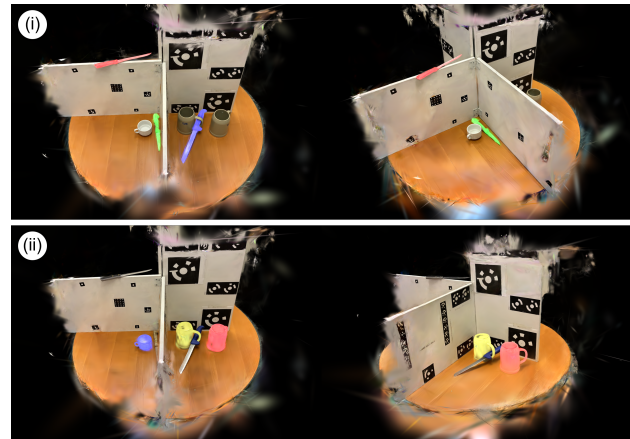


Figure 4. Qualitative results for Dataset 3. (i) Prompt: "Knife" & (ii) Prompt: "Mugs". Colors indicate recovered global instance identities after multiview association for both object categories

6. CONCLUSION

We present *TRACE*, a training-free framework for instance-level open vocabulary inventory generation in 3D crime scene reconstruction. The central contribution of the proposed method is a geometry- and appearance-aware multiview-consistent instancing formulation that assigns globally stable identities to objects of the same class. Building on these global instances, *TRACE* further allows for open-vocabulary scene querying by aggregating and lifting language aligned SigLIP2 descriptors into 3D Gaussian Splats.

The quantitative results show that geometry alone can be sufficient under strong inter-view connectivity, but also that appearance becomes essential once geometric continuity is weakened by structured occlusions. Across all three datasets, the combined formulation yields the most robust overall behaviour, while the qualitative lifting results demonstrate that the instance structure remains sufficiently stable to enable coherent instance aware inventory generation that can be directly addressed in 3D through natural language.

Overall, *TRACE* provides a step from semantic scene querying toward structured object-level access in reconstructed forensic scenes. Future work will investigate more robust prompt-based segmentation front-ends, broader evaluations on larger and more diverse forensic scenes and stronger quantitative assessment of the downstream 3D retrieval stage.

Table 2. Quantitative evaluation on Dataset 3. The scene contains five knives and four mugs under strong structured occlusions induced by cross-shaped partitions. In addition to the global clustering scores, class-wise ARI and B³ F1 values are reported for knives and mugs.

Method	Backbone	Global			Knife		Mug		#Pred/#GT
		ARI	B ³ F1	MVA	ARI	B ³ F1	ARI	B ³ F1	
Combined	DINOv2-B	0.935	0.937	1.000	0.853	0.875	1.000	1.000	12/9
	DINOv3-B	0.935	0.937	1.000	0.853	0.875	1.000	1.000	12/9
Geometry-only	–	0.904	0.906	1.000	0.827	0.859	0.956	0.954	14/9
Appearance-only	DINOv2-B	0.840	0.875	0.915	0.900	0.917	0.732	0.817	12/9
	DINOv3-B	0.864	0.909	0.906	1.000	1.000	0.682	0.805	10/9

References

- Balbudhe, M. S., Shetty, B. S. K., Arun Kumar, N., Singh, D., Sherekar, A. I., 2025. Crime scene reconstruction: a scoping review. *Egyptian Journal of Forensic Sciences*, 15(1), 79.
- Carew, R. M., French, J., Morgan, R. M., 2021. 3D forensic science: A new field integrating 3D imaging and 3D printing in crime reconstruction. *Forensic Science International: Synergy*, 3, 100205.
- Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A. et al., 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.
- Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q., 2025. Segment any 3d gaussians. *Proceedings of the AAAI conference on artificial intelligence*, 39(2), 1971–1979.
- Cheng, J., Zaech, J.-N., Van Gool, L., Pani Paudel, D., 2024. Occam’s LGS: A Simple Approach for Language Gaussian Splatting. *arXiv e-prints*.
- Halacheva, A.-M., Zaech, J.-N., Wang, X., Paudel, D. P., Van Gool, L., 2025. GaussianVLM: Scene-centric 3D Vision-Language Models using Language-aligned Gaussian Splats for Embodied Reasoning and Beyond. *arXiv preprint arXiv:2507.00886*.
- Jun-Seong, K., Kim, G., Yu-Ji, K., Wang, Y.-C. F., Choe, J., Oh, T.-H., 2025. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14137–14146.
- Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G. et al., 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.
- Kerr, J., Kim, C. M., Goldberg, K., Kanazawa, A., Tancik, M., 2023. Lrf: Language embedded radiance fields. *Proceedings of the IEEE/CVF international conference on computer vision*, 19729–19739.
- Li, Y., Ma, Q., Yang, R., Li, H., Ma, M., Ren, B., Popovic, N., Sebe, N., Konukoglu, E., Gevers, T. et al., 2025. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4961–4972.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., Chen, C., Wu, Z., 2024. 3d vision-language gaussian splatting. *arXiv preprint arXiv:2410.07577*.
- Piekenbrinck, J., Schmidt, C., Hermans, A., Vaskevicius, N., Linder, T., Leibe, B., 2025. Opensplat3d: Open-vocabulary 3d instance segmentation using gaussian splatting. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5246–5255.
- Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H., 2024. Langsplat: 3d language gaussian splatting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20051–20060.
- Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Schönberger, J. L., Zheng, E., Pollefeys, M., Frahm, J.-M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*.
- Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M. et al., 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.
- Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B. et al., 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- Wang, S., Li, K., Liang, S., Alegret, E., Ma, J., Navab, N., Gasperini, S., 2025. Visibility-Aware Language Aggregation for Open-Vocabulary Segmentation in 3D Gaussian Splatting. *arXiv preprint arXiv:2509.05515*.
- Ye, M., Danelljan, M., Yu, F., Ke, L., 2024. Gaussian grouping: Segment and edit anything in 3d scenes. *European conference on computer vision*, Springer, 162–179.

A. APPENDIX

Table 3. Full quantitative evaluation on Dataset 1 (eight distinct knives).

Method	ARI	B ³ Prec.	B ³ Rec.	B ³ F1	Hom.	Comp.	V-meas.	MVA	#Pred/#GT
Combined	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	8/8
Geometry-only	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	8/8
Appearance-only (DINOv2-B)	0.9265	0.9740	0.8847	0.9272	0.9809	0.8927	0.9347	0.9779	15/8
Appearance-only (DINOv3-B)	0.9872	0.9963	0.9786	0.9874	0.9975	0.9766	0.9870	0.9963	10/8

Table 4. Full quantitative evaluation on Dataset 2 (four distinct and four near-identical knives).

Method	ARI	B ³ Prec.	B ³ Rec.	B ³ F1	Hom.	Comp.	V-meas.	MVA	#Pred/#GT
Combined	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	8/8
Geometry-only	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	8/8
Appearance-only (DINOv2-B)	0.6477	0.7614	0.6064	0.6751	0.8167	0.6765	0.7400	0.8125	18/8
Appearance-only (DINOv3-B)	0.7503	0.8174	0.7395	0.7765	0.8574	0.7907	0.8227	0.8555	12/8

Table 5. Full global quantitative evaluation on Dataset 3 (five knives and four mugs under structured occlusion).

Setting	Backbone	ARI	B ³ Prec.	B ³ Rec.	B ³ F1	Hom.	Comp.	V-meas.	MVA	#Pred/#GT
Combined	DINOv2-B	0.9351	1.0000	0.8818	0.9372	1.0000	0.9259	0.9616	1.0000	12/9
	DINOv3-B	0.9351	1.0000	0.8818	0.9372	1.0000	0.9259	0.9616	1.0000	12/9
Geometry-only	–	0.9037	1.0000	0.8277	0.9057	1.0000	0.8923	0.9431	1.0000	14/9
Appearance-only	DINOv2-B	0.8398	0.8906	0.8597	0.8749	0.9249	0.8956	0.9100	0.9145	12/9
	DINOv3-B	0.8643	0.8822	0.9384	0.9094	0.9197	0.9470	0.9332	0.9060	10/9

Table 6. Full class-wise quantitative evaluation on Dataset 3.

Setting	Backbone	Class	ARI	B ³ Prec.	B ³ Rec.	B ³ F1	Hom.	Comp.	V-meas.	MVA	#Pred/#GT
Combined	DINOv2-B	Knife	0.8528	1.0000	0.7769	0.8745	1.0000	0.8281	0.9060	1.0000	8/5
		Mug	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	4/4
	DINOv3-B	Knife	0.8528	1.0000	0.7769	0.8745	1.0000	0.8281	0.9060	1.0000	8/5
		Mug	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	4/4
Geometry-only	–	Knife	0.8270	1.0000	0.7523	0.8586	1.0000	0.8036	0.8911	1.0000	9/5
		Mug	0.9556	1.0000	0.9127	0.9544	1.0000	0.9177	0.9571	1.0000	5/4
Appearance-only	DINOv2-B	Knife	0.9003	1.0000	0.8471	0.9172	1.0000	0.8672	0.9289	1.0000	8/5
		Mug	0.7323	0.7673	0.8739	0.8171	0.7463	0.8189	0.7809	0.8182	4/4
	DINOv3-B	Knife	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	5/5
		Mug	0.6821	0.7494	0.8691	0.8048	0.7286	0.8073	0.7659	0.8000	5/4