4DSUPERG: IDENTITY-AWARE SUPER-GAUSSIANS FOR OPEN-VOCABULARY 4D SCENE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary understanding of dynamic scenes requires text-selected objects to retain their identities as they move and become temporarily invisible. Existing 4D language fields distill per-Gaussian features through alpha-blended rendering, leaving primitive-level semantics ambiguous and often requiring feature compression. Identity-based language association addresses these issues in static scenes, but dynamic querying also requires the selected primitives to move coherently. We introduce 4DSuperG, which groups Gaussians into motion-coherent, identity-aware Super-Gaussians (SuperGs), making identity learning and motion prediction share the same unit. Each SuperG retrieves a full-dimensional CLIP feature through its object identity, without language-feature distillation. A query selects SuperGs once in canonical space, and their motion carries the selection through time and under occlusion. On HyperNeRF and Neu3D, 4DSuperG achieves state-of-the-art primitive-level open-vocabulary segmentation and competitive pixel-level performance, and supports zero-shot language-driven spatio-temporal editing.

1 INTRODUCTION

What should a representation for dynamic scene understanding offer? 4D Gaussian Splatting (4DGS) renders dynamic scenes with high fidelity (Luiten et al., 2024; Wu et al., 2024a), but applications such as mixed reality (Jiang et al., 2024) and interactive content creation (Mou et al., 2024; Shao et al., 2024) require more. Objects must be identifiable through free-form text queries, remain trackable over time, and retain their identities through occlusion, as shown in Figure 1. While existing 4D language fields answer such queries separately for each rendered view and only within visible regions, this paper presents a 4D representation that grounds each open-vocabulary query once, directly in scene primitives that carry object identity and persist regardless of per-frame visibility.

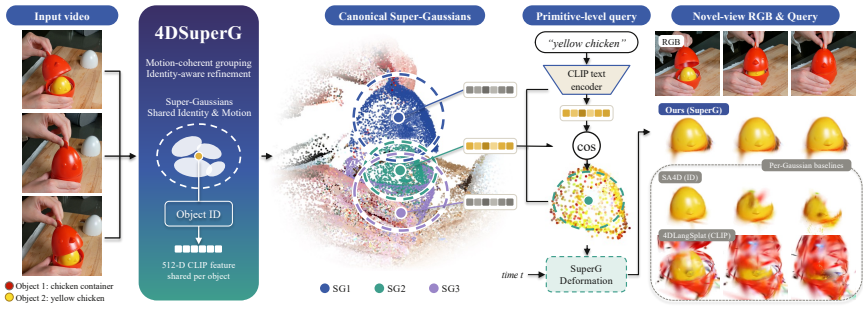

Figure 1: **4DSuperG grounds primitive-level open-vocabulary queries in motion-coherent, identity-aware Super-Gaussians.** SG1 and SG3 share the container's identity but follow different motions, while SG2 represents different object. Each SuperG's decoded object ID retrieves a full-dimensional CLIP feature shared by that object. A text query selects SuperGs once in canonical space, the selected primitives are then deformed and rendered across time. The chicken remains retrievable even when fully occluded by the container in the RGB view.

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.

Prior 4D language fields yield fragmented, semantically inconsistent selections when queried at the primitive level. These fields distill semantic features from 2D foundation models such as CLIP (Radford et al., 2021) by alpha-blending per-Gaussian language features into a rendered map and fitting it to 2D targets (Labe et al., 2024; Li et al., 2025; Zhou et al., 2025). However, alpha-blending constrains the rendered feature map without uniquely determining the features of individual Gaussians: multiple Gaussians contribute to each pixel, and different feature assignments satisfy the same supervision (Wu et al., 2024b). Accurate rendered features therefore do not guarantee accurate primitive-level features. Moreover, storing full-dimensional CLIP features per Gaussian is costly, so prior methods compress them to lower dimensions (Qin et al., 2024; Li et al., 2025), sacrificing fine-grained semantic distinctions (Zhou et al., 2024). Both issues are avoided by storing full-dimensional semantic features once per object identity and associating them with scene primitives without rendering-based optimization.

Optimization-free language association through object identities has been explored in static 3D scene understanding (Wu et al., 2024b). These methods use rendering to supervise a compact identity feature field that groups Gaussians into objects (Ye et al., 2024), then associate each object directly with a full-dimensional CLIP feature. This simplifies the learning task from *recovering continuous high-dimensional language features* to *distinguishing discrete identities*: identity feature errors do not change the label as long as the feature remains within the correct decision region, while observations of a static Gaussian across views often reinforce its identity. In dynamic scenarios, however, identifying the correct canonical primitives is only part of the problem: their trajectories must also preserve the selected object's structure over time. As shown in Figure 1, querying SA4D (Ji et al., 2024) equipped with the same identity-based association in canonical space yields a selection that fragments and distorts as its Gaussians deform across frames. Reliable primitive-level querying therefore requires consistency between object identity and motion.

Such consistency calls for a carrier larger than a single Gaussian. A dynamic scene contains far fewer distinct motion and identity units than rendering primitives. Predicting identity and motion per group therefore reduces the degrees of freedom constrained by limited 2D supervision and pools evidence across members and time. Visible members constrain the shared identity and motion, which also govern occluded members; during complete occlusion, the group retains its learned identity and follows its motion model. Forming these groups requires both signals: motion alone does not distinguish co-moving objects, while identity alone does not distinguish differently moving regions of the same object. Motion coherence guides grouping (Wan et al., 2024), while identity supervision discourages crossing object boundaries (Ye et al., 2024). 4DSuperG introduces this carrier as **identity-aware Super-Gaussians** (SuperGs), each grouping Gaussians that share an object identity and a motion model. Multiple SuperGs may share an identity while following different motions. Each SuperG decodes its object identity and retrieves a full-dimensional CLIP feature by lookup, extending identity-based language association to 4D. A query selects SuperGs once in canonical space, and their motion carries the selection through time. This structure also supports language-driven spatio-temporal editing, demonstrated qualitatively. *In 4DSuperG, the unit of identity learning is also the unit of motion.* We summarize our main contributions as follows.

- We introduce **identity-aware Super-Gaussians (SuperGs)**, a grouping learned jointly from motion and identity during reconstruction, as a carrier whose identity and trajectory stay consistent under occlusion and sparse supervision.
- We extend optimization-free, identity-based language association from static to dynamic Gaussians, attaching full-dimensional CLIP features without alpha-blending or dimensional compression and answering each query with a single evaluation in canonical space.
- We achieve state-of-the-art primitive-level open-vocabulary segmentation on HyperNeRF and Neu3D, on par with the strongest baseline at the pixel level, and attribute the gains through ablations: identity-based association over feature distillation, and the SuperG carrier over per-Gaussian identity. We further demonstrate zero-shot language-driven spatio-temporal editing qualitatively.

2 Related Works

Open-Vocabulary Understanding. Augmenting 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) with full-dimensional language features requires substantial costs. Feature 3DGS, LangSplat, and

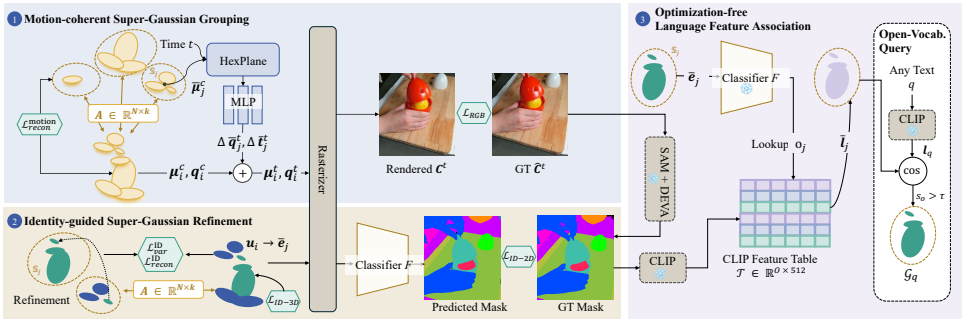

Figure 2: **Overview of 4DSuperG. (1)** The association matrix $\boldsymbol{A}$ groups canonical Gaussians into motion-coherent SuperGs, each with a shared rigid transformation. **(2)** SAM+DEVA masks supervise per-Gaussian encodings $\boldsymbol{u}_i$ to bootstrap shared SuperG identities $\bar{\boldsymbol{e}}_j$, which are then directly supervised to refine grouping along object boundaries. **(3)** Decoded IDs o_j retrieve full-dimensional CLIP features by lookup, enabling open-vocabulary queries without language-feature optimization.

LEGaussians reduce these costs through compact features or quantization, learned with rendering-based supervision (Zhou et al., 2024; Qin et al., 2024; Shi et al., 2024). Instance-level sharing offers another solution: OpenGaussian associates instances with CLIP features (Wu et al., 2024b), whereas subsequent work learns instance and language codebooks linked by attention (Alegret et al., 2026), or instead avoid rendered-feature optimization by directly aggregating multi-view features onto Gaussians (Jun-Seong et al., 2025; Wang et al., 2026). These approaches address static scenes.

Dynamic language fields such as DGD (Labe et al., 2024) and Feature4X (Zhou et al., 2025) extend rendering-based feature distillation to moving Gaussians. To capture temporal semantics, 4-LEGS (Fiebelman et al., 2025) distills video-language features, while 4DLangSplat (Li et al., 2025) uses embeddings of MLLM-generated captions. DLGStream (Ke et al., 2026) separates color and language opacities to support streaming. These methods retain rendering-based language supervision, which constrains individual Gaussian semantics indirectly. Recent methods incorporate instance information per-Gaussian: LangField4D conditions language features on adaptive instance labels (Xu et al., 2026), while ST4R-Splat combines referring embeddings with instance-level temporal descriptors (Meng et al., 2026). 4DSuperG instead grounds queries in shared SuperG that linked to full-dimensional language features, with each group also being a coherent motion unit.

Instance Segmentation. Instance segmentation in 3DGS transfers supervision from 2D models such as SAM (Kirillov et al., 2023). Gaussian Grouping (Ye et al., 2024) learns per-Gaussian ID encoding by classifying rendered features against 2D masks. SA4D (Ji et al., 2024) introduces a temporal identity field for dynamic scenes, while TRASE (Li et al., 2026) uses contrastive supervision to avoid requiring temporally consistent masks. Segment–then–Splat (Lu et al., 2025) fixes Gaussian–object membership before reconstruction. Relatively, 4DSuperG jointly constrains grouping by motion and identity, supervising shared SuperG identities to reduce sensitivity to per-Gaussian noise.

Superpoints and Super-Gaussians. Grouping points into structured primitives supports efficient scene understanding (Robert et al., 2023; 2024). SuperGSeg (Liang et al., 2026) combines geometric and segmentation cues to form SuperGs, then distills language features onto these groups for compact open-vocabulary understanding of static scenes. For dynamic reconstruction, SP-GS (Wan et al., 2024) learns superpoints with an ARAP-inspired objective (Sorkine & Alexa, 2007) and predicts shared rigid transformations to accelerate novel-view synthesis. However, motion coherence does not ensure a common object identity, since distinct objects may move together. 4DSuperG complements motion-based grouping with ID encoding supervision, encouraging SuperGs to preserve object boundaries while maintaining coherent motion.

3 METHOD

The 4DSuperG pipeline is visualized in Figure 2. The 4D Gaussians are first reconstructed and grouped into SuperGs using RGB supervision and motion coherence (Section 3.1), followed by

identity-guided refinement under continued motion supervision (Section 3.2). A mapping from per-SuperG object IDs to full-dimensional CLIP features enables compact storage and direct primitive-level querying (Section 3.3).

3.1 MOTION-COHERENT SUPER-GAUSSIAN GROUPING

Per-Gaussian deformation predicts motion separately for each primitive, so a selected object fragments and distorts as its Gaussians deform across frames, as shown in Figure 3. 4DSuperG therefore builds on SP-GS (Wan et al., 2024) to group Gaussians by motion coherence and predict shared motion per SuperG. Throughout, $\bar{(\cdot)}$ denotes SuperG-level attributes, subscripts i and j index Gaussians and SuperGs, and superscripts c and t denote canonical and time-t states, respectively.

Gaussian-to-SuperG association. 4DSuperG initializes M SuperGs $\{\mathbb{S}_j\}_{j=1}^{M}$ centers by farthest point sampling (Qi et al., 2017) on the canonical Gaussian positions. Each of the N Gaussians $\mathcal{G}_i$ keeps its k nearest initial centers as a candidate set $\mathcal{N}_i^k$ with learnable logits ℓ_{ij}. Association weights are a softmax over the candidates, and each SuperG center is recomputed at every iteration as an association-weighted centroid, with gradients stopped:

$$a_{ij} = \mathbb{1}[j \in \mathcal{N}_i^k] \frac{\exp(\ell_{ij})}{\sum_{j' \in \mathcal{N}_i^k} \exp(\ell_{ij'})}, \qquad \bar{\boldsymbol{\mu}}_j^c = \frac{\sum_{i:\, j \in \mathcal{N}_i^k} a_{ij} \boldsymbol{\mu}_i^c}{\sum_{i:\, j \in \mathcal{N}_i^k} a_{ij}}, \tag{1}$$

The weights form a sparse matrix $\boldsymbol{A} \in \mathbb{R}^{N \times k}$ whose i-th row stores a_{ij} over $\mathcal{N}_i^k$ and the centroid update keeps the centers on the evolving canonical geometry.

SuperG–based Deformation Network. Next, 4DSuperG predicts a shared motion for each SuperG. A HexPlane encoder (Cao & Johnson, 2023) and two MLP decoders map each SuperG's canonical center and time $(\bar{\boldsymbol{\mu}}_j^c, t)$ to a translation $\Delta\bar{\boldsymbol{t}}_j^t$ and unit quaternion $\Delta\bar{\boldsymbol{q}}_j^t$. Each candidate SuperG rotates Gaussian $\mathcal{G}_i$ about the SuperG's canonical center and translates it:

$$\boldsymbol{\mu}_{i\leftarrow j}^t = \underbrace{(\bar{\boldsymbol{\mu}}_j^c + \Delta\bar{\boldsymbol{t}}_j^t)}_{\text{translated SuperG center}} + \underbrace{\mathcal{R}(\Delta\bar{\boldsymbol{q}}_j^t)}_{\text{rotation}} \underbrace{(\boldsymbol{\mu}_i^c - \bar{\boldsymbol{\mu}}_j^c)}_{\text{canonical offset}}, \quad \boldsymbol{q}_{i\leftarrow j}^t = \Delta\bar{\boldsymbol{q}}_j^t \otimes \boldsymbol{q}_i^c, \tag{2}$$

where $\otimes$ denotes quaternion composition and $\mathcal{R}$ converts a unit quaternion to a rotation matrix. Each Gaussian blends its candidate positions and rotations with the association weights:

$$\boldsymbol{\mu}_i^t = \sum_{j \in \mathcal{N}_i^k} a_{ij} \boldsymbol{\mu}_{i\leftarrow j}^t, \qquad \boldsymbol{q}_i^t = \text{normalize}\left(\sum_{j \in \mathcal{N}_i^k} a_{ij} \boldsymbol{q}_{i\leftarrow j}^t \right), \tag{3}$$

where candidate quaternions are sign-aligned before blending.

The deformed Gaussians are alpha-composited into an image $\boldsymbol{C}^t$, and the photometric loss $\mathcal{L}_{RGB} = \|\boldsymbol{C}^t - \hat{\boldsymbol{C}}^t\|_1$ against the ground-truth image $\hat{\boldsymbol{C}}^t$ jointly optimizes the association logits, the deformation network, and the Gaussian attributes.

An optional per-Gaussian non-rigid residual, adapted from SP-GS (Wan et al., 2024), captures fine-scale deformation and is trained once SuperG optimization is complete, with all other parameters frozen (Appendix A.4).

Motion-Aware Association Learning. To favor candidates whose motion agrees with each Gaussian's highest-weight SuperG, $j_i^* = \arg\max_{j \in \mathcal{N}_i^k} a_{ij}$, we introduce a motion-consistency loss. Let candidate displacement be $\Delta\boldsymbol{\mu}_{i\leftarrow j}^t = \boldsymbol{\mu}_{i\leftarrow j}^t - \boldsymbol{\mu}_i^c$ and relative orientation be $\Delta\boldsymbol{q}_{i\leftarrow j}^t = \Delta\bar{\boldsymbol{q}}_j^t$:

$$\mathcal{L}_{recon}^{\text{motion}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i^k} a_{ij} \Big[\|\Delta\boldsymbol{\mu}_{i\leftarrow j}^t - \Delta\boldsymbol{\mu}_{i\leftarrow j_i^*}^t\|_2^2 + \lambda_1 \left(1 - \left[(\Delta\boldsymbol{q}_{i\leftarrow j}^t)^\top \Delta\boldsymbol{q}_{i\leftarrow j_i^*}^t \right]^2 \right) \Big]. \tag{4}$$

All candidate displacements and rotations in this loss are detached, so it updates only the association logits, suppressing candidates whose motion disagrees with the primary SuperG. In short, the overall objective for motion-coherent SuperG grouping is: $\mathcal{L}_{\text{motion}} = \mathcal{L}_{RGB} + \lambda_{recon}^{\text{motion}} \mathcal{L}_{recon}^{\text{motion}}$.

3.2 IDENTITY-GUIDED SUPER-GAUSSIAN REFINEMENT

Motion alone does not distinguish static or co-moving objects, so motion-based grouping may cross object boundaries, as shown in Figure 6. To separate these objects, each SuperG carries a learnable identity encoding $\bar{\boldsymbol{e}}_j$, rendered and classified per pixel under tracked 2D mask supervision (Ye et al., 2024; Ji et al., 2024). This supervision refines the associations learned in Section 3.1, encouraging each SuperG to group Gaussians from the same object.

Per-Gaussian identity bootstrapping. Shared identity learning and association refinement are interdependent: *object-consistent groups facilitate learning discriminative shared encodings*, while *refining the associations requires discriminative identity cues*. We therefore first learn temporary per-Gaussian encodings $\boldsymbol{u}_i \in \mathbb{R}^D$ to refine the associations and initialize the shared encodings.

At time t, the unit-normalized per-Gaussian encodings are rendered at the deformed positions (Equation 3) into an identity feature map $\boldsymbol{E}^t \in \mathbb{R}^{H\times W\times D}$. A linear classifier F followed by softmax maps each pixel feature to probabilities over O object labels. Supervision comes from SAM masks (Kirillov et al., 2023; Zrporz, 2024) tracked across frames with DEVA (Cheng et al., 2023). The encodings and classifier are optimized with pixel-wise cross-entropy $\mathcal{L}_{ID-2D}$ and the 3D regularizer $\mathcal{L}_{ID-3D}$ (Ye et al., 2024; Ji et al., 2024). Preprocessing and loss details are given in Appendix A.

The per-Gaussian encodings refine the associations through an identity variance loss that compares each Gaussian's encoding with its candidates' association-weighted mean encodings:

$$\mathcal{L}_{var}^{\mathrm{ID}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i^k} a_{ij}\left\|\boldsymbol{u}_i - \bar{\boldsymbol{u}}_j\right\|_2^2, \qquad \bar{\boldsymbol{u}}_j = \frac{\sum_{i:\, j\in\mathcal{N}_i^k} a_{ij}\boldsymbol{u}_i}{\sum_{i:\, j\in\mathcal{N}_i^k} a_{ij}}. \tag{5}$$

$\boldsymbol{u}_i$ are detached throughout this computation, so the loss updates only the association logits, favoring candidates whose mean encodings better match each Gaussian's identity.

The bootstrap objective is:

$$\mathcal{L}_{\mathrm{bootstrap}} = \mathcal{L}_{\mathrm{motion}} + \lambda_{\mathrm{2D}}\mathcal{L}_{ID-2D} + \lambda_{\mathrm{3D}}\mathcal{L}_{ID-3D} + \lambda_{\mathrm{var}}\mathcal{L}_{var}^{\mathrm{ID}}. \tag{6}$$

The motion objective remains active during bootstrapping, encouraging groups that are coherent in both motion and identity.

Shared SuperG identity refinement. We replace the temporary encodings with one learnable, time-invariant encoding $\bar{\boldsymbol{e}}_j$ per SuperG, initialized from $\bar{\boldsymbol{u}}_j$. Unlike independent per-Gaussian encodings, shared encodings allow mask supervision to refine the associations directly.

We render $\boldsymbol{E}^t$ using the shared encodings through the Gaussian–SuperG associations and reuse F, $\mathcal{L}_{\mathrm{ID-2D}}$, and $\mathcal{L}_{\mathrm{ID-3D}}$. When a SuperG spans different objects, conflicting mask supervision encourages associations with candidates whose encodings better explain the object labels. Visible members thus update encodings also used by occluded members. To encourage identity consistency among each Gaussian's candidates, we use the identity reconstruction loss:

$$\mathcal{L}_{recon}^{\mathrm{ID}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\mathcal{N}_i^k} a_{ij}\left(1 - \bar{\boldsymbol{e}}_j^{\top}\bar{\boldsymbol{e}}_{j_i^*}\right). \tag{7}$$

With both encodings detached, this loss updates only association logits, favoring identity agreement with the primary SuperG. Shared refinement replaces $\mathcal{L}_{var}^{\mathrm{ID}}$ in Equation 6 with $\mathcal{L}_{recon}^{\mathrm{ID}}$, retaining motion supervision. For inference, notably, we use only the learned shared encodings per SuperG.

3.3 OPTIMIZATION-FREE LANGUAGE FEATURE ASSOCIATION

The refined SuperGs carry object identities that connect scene primitives directly to full-dimensional 2D CLIP features. This identity-based lookup requires no language-feature optimization or rendering, avoiding both the ambiguous primitive-level supervision induced by alpha-blending (Wu et al., 2024b) and the feature compression commonly used in per-Gaussian language distillation (Li et al., 2025).

Identity decoding and feature lookup. The classifier F from Section 3.2 decodes each SuperG's object label once from its normalized shared encoding:$o_j = \arg\max_{o\in\{1,\ldots,O\}} F(\bar{\boldsymbol{e}}_j)_o$. Using the



Figure 3: **Qualitative comparison of 4D open-vocabulary object selection.** Results on (a) HyperNeRF *Torchocolate*, (b) Neu3D *Sear Steak*, (c) HyperNeRF *Chickchicken*, and (d) Neu3D *Cut Roasted Beef* across two views. For each scene, we show RGB observations and primitive-level selections by our method, SA4D, and 4DLangSplat for the listed language queries.

tracked masks from identity supervision, we encode each object's masked crops with the CLIP image encoder and average the embeddings across observed frames to obtain the time-invariant table $\mathcal{T}_{\text{CLIP}}(o) \in \mathbb{R}^{512}$. Each SuperG accesses its language feature by lookup: $\bar{\boldsymbol{l}}_j = \mathcal{T}_{\text{CLIP}}(o_j)$.

Open-Vocabulary Querying in Canonical Space. With language features associated through object identities, an open-vocabulary query q can be resolved directly in canonical space and propagated across time. Given a text query q, the CLIP text encoder produces $\boldsymbol{l}_q \in \mathbb{R}^{512}$. We compute one cosine similarity per object ID instead of per Gaussian, select SuperGs whose scores exceed a threshold τ, and retrieve Gaussians through their primary associations:

$$s_o = \cos(\mathcal{T}_{\text{CLIP}}(o), \boldsymbol{l}_q), \quad \mathcal{J}_q = \{j \mid s_{o_j} > \tau\}, \quad \mathcal{G}_q = \{\mathcal{G}_i \mid j_i^* \in \mathcal{J}_q\}. \tag{8}$$

An optional identity-aware post-pruning removes Gaussians whose inherited object identity conflicts with tracked mask evidence, yielding $\mathcal{G}_q^*$ (Appendix A.3). The pruning scores are computed once offline and are independent of the query.

The selected Gaussians keep their SuperG deformation and can be rendered at every time step. Language-driven editing follows the same path: a query selects Gaussians once in canonical space where the edit is applied, and the SuperG deformation carries it to every time step, so the edit stays consistent across views and time without per-frame processing.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior work on 4D open-vocabulary understanding, we evaluate on HyperNeRF (Park et al., 2021) and Neu3D (Li et al., 2022). HyperNeRF provides monocular sequences with non-rigid motion and topological changes, while Neu3D captures multi-object interactions using a synchronized multi-camera rig. We evaluate six scenes per dataset using ground-truth segmentation masks and text queries from 4DLangSplat (Li et al., 2025).

Table 1: **Open-vocabulary 4D segmentation on HyperNeRF and Neu3D.** Scores are averaged across scenes and reported in percent. Colors mark **first**, second, and third place. * denotes compressed language features, † denotes ID-to-CLIP lookup, and $^\diamond$ denotes post-pruning for primitive-level querying. ✗ indicate unsupported results. Per-scene results are in the Appendix.

Method	*HyperNeRF*				*Neu3D*			
	Pixel-level		Primitive-level		Pixel-level		Primitive-level	
	mIoU↑	mAcc@50↑	mIoU↑	mAcc@50↑	mIoU↑	mAcc@50↑	mIoU↑	mAcc@50↑
DGD* (Labe et al., 2024)	62.24	67.75	40.08	47.48	✗	✗	✗	✗
4DLangSplat* (Li et al., 2025)	85.06	93.86	52.61	63.60	59.50	64.32	51.89	60.42
Feature4X* (Zhou et al., 2025)	57.81	61.28	43.62	48.77	✗	✗	✗	✗
DLGStream* (Ke et al., 2026)	✗	✗	✗	✗	78.36	90.85	17.48	12.23
TRASE (Li et al., 2026)	✗	✗	48.56	53.21	✗	✗	24.39	16.14
SA4D†* (Ji et al., 2024)	84.26	91.44	62.86	73.88	75.48	80.00	52.67	59.46
SA4D† (Ji et al., 2024)	84.58	91.90	63.06	74.72	83.52	**93.33**	59.41	66.47
SA4D$^{\dagger\diamond\star}$ (Ji et al., 2024)	84.26	91.44	71.25	83.54	75.48	80.00	69.55	84.71
SA4D$^{\dagger\diamond}$ (Ji et al., 2024)	84.58	91.90	71.94	85.75	83.52	**93.33**	73.64	91.37
Ours$^{\dagger\diamond\star}$	82.86	90.38	76.55	89.02	80.63	86.67	72.88	86.67
Ours$^{\dagger\diamond}$	86.25	94.14	79.10	92.81	**84.24**	**93.33**	76.07	**93.33**
Ours$^{\dagger\diamond}$ **+ non-rigid**	**86.89**	**94.64**	**79.22**	**93.26**	84.11	**93.33**	**77.37**	**93.33**

Baselines. We compare against baselines that encode either language features or object identities per-Gaussian. DGD (Labe et al., 2024), 4DLangSplat (Li et al., 2025), and Feature4X (Zhou et al., 2025) distill per-Gaussian language features through alpha-blended feature rendering and answer queries by matching rendered features to text embeddings. Notably, DGD and Feature4X support only monocular input. TRASE (Li et al., 2026) and SA4D (Ji et al., 2024) instead encode per-Gaussian identity without storing language features. TRASE retrieves Gaussian clusters using Grounded SAM (Ren et al., 2024) on rendered views. It lacks rendered language features for pixel-wise matching and is therefore evaluated only under the primitive-level querying protocol. For SA4D, we equip its per-Gaussian identity field with the proposed CLIP features lookup Section 3.3, so the comparison isolates the carrier, per-Gaussian against SuperG, under the same language association.

Metrics and protocols. All methods are evaluated under two protocols where applicable. The **pixel-level protocol** resolves each query per rendered view. The **primitive-level protocol** selects primitives once in canonical space and renders the selected subset at every time step. Both report mIoU and mAcc@50 (the fraction of queries with IoU above 0.5) against 2D ground-truth masks. The pixel-level score measures segmentation in the rendered image, where blending can hide errors of individual primitives. Post-pruning is applied only to primitive-level querying and therefore does not alter pixel-level scores. The primitive-level score measures whether the primitives themselves carry the right semantics and keep it persistently under deformation.

Implementation details. 4DSuperG is trained on one NVIDIA RTX 4090 with Adam (Kingma & Ba, 2015) for 23,000 iterations per scene (same as 4DLangSplat). All experiments use $M = 600$ SuperGs, $k = 3$ candidates, and 512-D CLIP descriptors. The stage schedule is detailed in Appendix B.4, data splits and fixed thresholds in Appendix B.2, and ablations over M and k in Appendix C.1.

4.2 Spatiotemporal Open-Vocabulary Segmentation

Table 1 compares pixel- and primitive-level open-vocabulary segmentation on HyperNeRF and Neu3D. Our method performs on par with the strongest baselines at the pixel level while achieving substantially higher primitive-level accuracy. The qualitative results in Figure 3 further show cleaner primitive selections and better preservation of object structure under occlusion.

Identity to 512D CLIP lookup improves primitive-level performance. Distillation-based 4D language fields (Labe et al., 2024; Li et al., 2025; Zhou et al., 2025; Ke et al., 2026) lose substantial accuracy under primitive-level evaluation. With compressed language features, 4DLangSplat* leads Ours* by 2.20 pixel-level mIoU on HyperNeRF, yet trails it by 23.94 at the primitive level. Fitting alpha-blended feature maps does not uniquely constrain individual Gaussian features, resulting in noisy, incomplete primitive selections (Figure 3). At the same 3D language descriptor, SA4D$^{\dagger\diamond\star}$, equipped with our identity-to-CLIP lookup and post-pruning, exceeds 4DLangSplat* by 18.64 and 17.66 primitive-level mIoU on HyperNeRF and Neu3D. Increasing descriptor dimensionality to 512

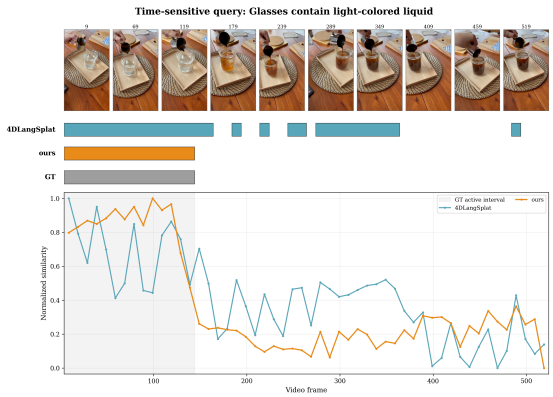

Figure 4: **Time-sensitive querying.** We compare 4DLangSplat (Li et al., 2025) with our method. The bottom row shows cosine similarity across frames, rescaled to (0,1). Horizontal bars mark the retrieved time segments.

Table 2: **Component ablations.** Primitive-level scores (%) averaged over six scenes per dataset. ⋆: compressed 3-D language features (otherwise 512-D). SG: motion-coherent SuperG grouping; IB: per-Gaussian identity bootstrap; SIR: shared SuperG identity refinement; DA: direct language association; PP: post-pruning; NR: non-rigid residual. mAcc denotes mAcc@50.

Case	SG	IB	SIR	DA	PP	NR	*HyperNeRF*		*Neu3D*	
							mIoU	mAcc	mIoU	mAcc
(a)⋆							52.61	63.60	51.89	60.42
(b)⋆	✓						61.51	75.33	61.66	70.78
(c)⋆	✓	✓					62.71	77.17	57.66	59.79
(d)⋆	✓	✓	✓	✓			70.29	86.13	68.20	80.20
(e)	✓	✓		✓			70.54	84.95	74.75	93.33
(f)	✓		✓	✓			67.27	82.27	50.79	63.11
(g)	✓	✓	✓	✓			73.23	89.43	75.31	**96.00**
(h)	✓	✓	✓	✓	✓		79.10	92.81	76.07	93.33
(i)	✓	✓	✓	✓	✓	✓	**79.22**	**93.26**	**77.37**	93.33

further improves primitive-level mIoU by 0.69 and 4.09 for SA4D$^{\dagger\diamond}$, and by 2.55 and 3.19 for Ours, on the two datasets, respectively. Identity lookup supports these full-dimensional descriptors through a shared object-level table, without optimizing or storing a 512D language feature for every Gaussian.

Super-Gaussians outperform per-Gaussian identities. With the same lookup and post-pruning, the comparison against SA4D$^{\dagger\diamond}$ tests the benefit of the SuperG carrier. With full-dimensional descriptors, Ours exceeds SA4D$^{\dagger\diamond}$ by 7.16 mIoU and 7.06 mAcc@50 on HyperNeRF, and by 2.43 mIoU and 1.96 mAcc@50 on Neu3D at the primitive level. The mIoU advantage remains with 3D descriptors: 5.30 on HyperNeRF and 3.33 on Neu3D. The larger gain on monocular HyperNeRF is consistent with shared identity learning being particularly useful under sparse viewpoint supervision. By pooling identity evidence within each motion-coherent SuperG, our method assigns its members a shared identity that persists when individual Gaussians are occluded, as shown in Figure 3.

Super-Gaussians preserve object structure through occlusion. In Figure 3, 4DLangSplat selects scattered floaters and streaks alongside the queried objects. SA4D with our augmentation reduces these artifacts but retains thin boundary streaks and fragments the yellow chicken in (c). Our SuperG-based selections have cleaner boundaries and present fewer artifacts across views. The sequence in Figure 1 shows each method deforming the same canonically selected Gaussians as the container closes. SA4D's initially coherent chicken becomes fragmented, while 4DLangSplat selects container and background Gaussians from the outset. Our selection retains the chicken's shape even when it is fully occluded in the RGB view, as its Gaussians follow shared SuperG motion.

4.3 Ablation Study

Table 2 examines how the main components contribute to primitive-level segmentation on six scenes per dataset. Motion-coherent SuperG grouping already improves over 4DLangSplat, (a)→(b), by 8.90 mIoU on HyperNeRF and 9.77 on Neu3D, showing that grouping benefits primitive selection even with distilled language features. However, adding identity bootstrap in (c) yields only a small gain on HyperNeRF and a drop on Neu3D, indicating identity supervision does not consistently translate into better language selection **when features are still distilled through rendering**. Combining shared identity refinement with direct language association in (d) then raises mIoU by 7.58 and 10.54, respectively, supporting the combination of shared identity learning and descriptor lookup. Moreover, retaining full 512-D descriptors, (d)→(g), adds another 2.94 and 7.11 mIoU, showing that compression discards information useful for primitive-level queries.

The identity-stage ablations further support bootstrapping before shared refinement. Without the per-Gaussian bootstrap, (g)→(f), mIoU falls by 5.96 on HyperNeRF and 24.52 on Neu3D, indicating that initial per-Gaussian identity supervision provides an important starting point for learning shared identities. Shared refinement also contributes: removing it, (g)→(e), reduces mIoU by 2.69 and

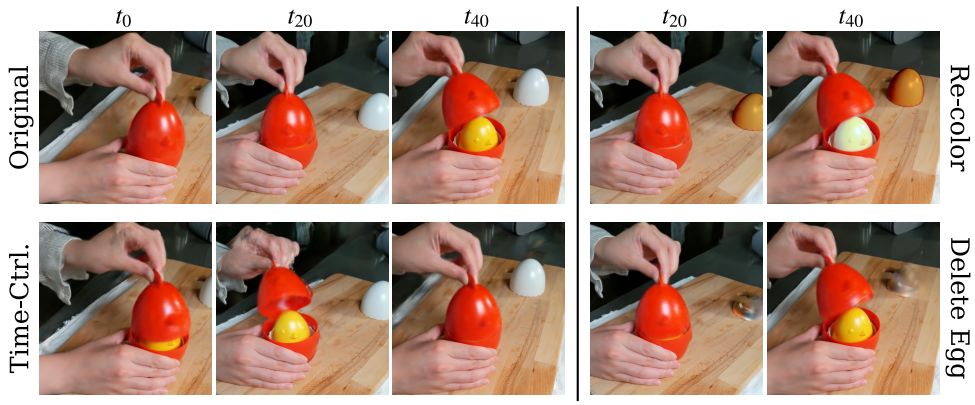

Figure 5: **Spatial-temporal consistent native 3D editing.** 4DSuperG allow to shuffle the motion in time while keeping the scene intact, and re-colors or deletes objects consistently across frames.

0.56, respectively. Together, these comparisons support the two-stage design, which first establishes identity evidence on individual Gaussians and then consolidates it within SuperGs.

Even without post-pruning or non-rigid residuals, case (g) exceeds SA4D† by 10.17 primitive-level mIoU on HyperNeRF and 15.90 on Neu3D under the same full-dimensional identity lookup. The two optional modules provide further gains: post-pruning, (g)→(h), adds 5.87 and 0.76 mIoU, while the non-rigid residual, (h)→(i), adds another 0.12 and 1.30, respectively.

4.4 Super-Gaussians as an Object-Level Interface

A reusable representation for object-level features. Once trained, each SuperG carries an object ID through which any object-level 2D feature can be attached without retraining. For time-sensitive queries, we replace the static CLIP table with a frame-indexed MLLM-caption feature table $\mathcal{T}_{\mathrm{MLLM}}(t, o)$; at time t, each pixel's decoded object ID selects the corresponding feature (Appendix B.3). Unlike 4DLangSplat (Li et al., 2025), which requires separate distillation, our direct association improves scene-macro vIoU and mAcc by 29.02 and 12.34 points on four HyperNeRF scenes (Figure 4); Appendix C.3 provides per-scene results.

Native object-level editing with temporal consistency. SuperGs also serve as 3D editing units, supporting object deletion, recoloring, and time control, as shown in Figure 5. Edits operate on objects selected once in canonical space and propagate through the SuperG motion model to every time step, preserving spatial and temporal consistency without frame-wise editing or further optimization.

5 Conclusion

We present 4DSuperG, a 4D representation that grounds open-vocabulary queries in identity-aware Super-Gaussians. By coupling object identity with coherent motion, 4DSuperG allows a query to select primitives once in canonical space and follow them through time, even during occlusion. Decoded object IDs index full-dimensional CLIP features, bypassing the ambiguous primitive-level supervision and compression of alpha-blended language-feature distillation. On HyperNeRF and Neu3D, 4DSuperG achieves state-of-the-art primitive-level segmentation and performs on par with the strongest baseline at the pixel level.

The learned identities also provide an interface for attaching object-level features from other 2D models without retraining the representation, as illustrated by MLLM features for time-sensitive queries. Canonical selection further enables zero-shot language-driven edits that remain consistent across frames. These capabilities make SuperGs promising object-level handles for downstream applications that need to refer to and manipulate the same entities over time, including mixed-reality interaction and interactive editing.

AI USE STATEMENT

Generative AI tools were used to assist with programming tasks and to improve the grammar, clarity, and presentation of the manuscript. They were not used for synthetic data generation, nor for mathematical formulation or proofs, both of which are not applicable to this work. All AI-assisted code and text were reviewed, revised, and validated by the authors. The research ideas, problem formulation, method development, experimental design, data analysis, and scientific conclusions were conceived and carried out by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available. For HyperNeRF, we strictly follow the pre-processing procedure released by 4DLangSplat (Li et al., 2025). For Neu3D, we build on the same procedure and additionally enforce consistency across its synchronized camera views. Each baseline is trained on the same processed data using its official open-source implementation and evaluated under our unified evaluation protocols, ensuring consistent data and metrics across methods. The appendix documents the training settings, baseline adaptations, and evaluation protocols. We will publicly release our implementation and complete evaluation code to facilitate reproduction of all reported results.

REFERENCES

Elena Alegret, Kunyi Li, Sen Wang, Siyun Liang, Michael Niemeyer, Stefano Gasperini, Nassir Navab, and Federico Tombari. GALA: guided attention with language alignment for open vocabulary gaussian splatting. In *3DV*. IEEE, 2026.

Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *CVPR*, pp. 130–141. IEEE, 2023.

Ho Kei Cheng, Seoung Wug Oh, Brian L. Price, Alexander G. Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, pp. 1316–1326. IEEE, 2023.

Gal Fiebelman, Tamir Cohen, Ayellet Morgenstern, Peter Hedman, and Hadar Averbuch-Elor. 4-legs: 4d language embedded gaussian splatting. *Comput. Graph. Forum*, 44(2), 2025.

Shengxiang Ji, Guanjun Wu, Jiemin Fang, Jiazhong Cen, Taoran Yi, Wenyu Liu, Qi Tian, and Xinggang Wang. Segment any 4d gaussians, 2024.

Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Y. K. Lau, Feng Gao, Yin Yang, and Chenfanfu Jiang. VR-GS: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *SIGGRAPH (Conference Paper Track)*, pp. 78. ACM, 2024.

Kim Jun-Seong, GeonU Kim, Kim Yu-Ji, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14137–14146. IEEE, 2025.

Zhihui Ke, Yuyang Liu, Xiaobo Zhou, and Tie Qiu. Dlgstream: Dynamic language-embedded gaussian splatting for open-vocabulary enabled free-viewpoint video streaming. In *European Conference on Computer Vision*, pp. 215–237. Springer, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Isaac Labe, Noam Issachar, Itai Lang, and Sagie Benaim. DGD: dynamic 3d gaussians distillation. In *ECCV*, pp. 361–378. Springer, 2024.

Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *CVPR*, pp. 5511–5521. IEEE, 2022.

Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 4d langsplat: 4d language gaussian splatting via multimodal large language models. In *CVPR*, pp. 22001–22011. IEEE, 2025.

Yun-Jin Li, Mariia Gladkova, Yan Xia, and Daniel Cremers. Trase: Tracking-free 4d segmentation and editing. In *3DV*. IEEE, 2026.

Siyun Liang, Sen Wang, Kunyi Li, Michael Niemeyer, Stefano Gasperini, Hendrik Lensch, Nassir Navab, and Federico Tombari. Supergseg: Open-vocabulary 3d segmentation with structured super-gaussians. In *3DV*. IEEE, 2026.

Yiren Lu, Yunlai Zhou, Yiran Qiao, Chaoda Song, Tuo Liang, Jing Ma, Huan Wang, and Yu Yin. Segment then splat: Unified 3d open-vocabulary segmentation via gaussian splatting. *Advances in Neural Information Processing Systems*, 38:165975–165994, 2025.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, pp. 800–809. IEEE, 2024.

Yuming Meng, Dong Wu, and Hongbin Zha. St4r-splat: Spatio-temporal referring segmentation in 4d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17598–17608, 2026.

Linzhan Mou, Jun-Kun Chen, and Yu-Xiong Wang. Instruct 4d-to-4d: Editing 4d scenes as pseudo-3d scenes using 2d diffusion. In *CVPR*, pp. 20176–20185. IEEE, 2024.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6):238:1–238:12, 2021.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pp. 5099–5108, 2017.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, pp. 20051–20060. IEEE, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Damien Robert, Hugo Raguet, and Loïc Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *ICCV*, pp. 17149–17158. IEEE, 2023.

Damien Robert, Hugo Raguet, and Loïc Landrieu. Scalable 3d panoptic segmentation as superpoint graph clustering. In *3DV*, pp. 179–189. IEEE, 2024.

Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Control4d: Efficient 4d portrait editing with text. In *CVPR*, pp. 4556–4567. IEEE, 2024.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *CVPR*, pp. 5333–5343. IEEE, 2024.

Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry Processing*, pp. 109–116. Eurographics Association, 2007.

Diwen Wan, Ruijie Lu, and Gang Zeng. Superpoint gaussian splatting for real-time high-fidelity dynamic scene reconstruction. In *ICML*, pp. 49957–49972. PMLR, 2024.

Sen Wang, Kunyi Li, Siyun Liang, Elena Alegret, Jing Ma, Nassir Navab, and Stefano Gasperini. Visibility-aware language aggregation for open-vocabulary segmentation in 3d gaussian splatting. In *3DV*. IEEE, 2026.

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 *CVPR*, pp. 20310–20320. IEEE, 2024a.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In *NeurIPS*, 2024b.

Yichao Xu, Qiaowei Miao, Jinsheng Quan, Wei Yang, Zhihui Li, and Yawei Luo. Langfield4d: Learning identity-adaptive and spatio-temporal continuous 4d language fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9558–9569, 2026.

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

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatiotemporal video grounding for multi-form sentences. In *CVPR*, pp. 10665–10674. Computer Vision Foundation / IEEE, 2020.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *CVPR*, pp. 21676–21685. IEEE, 2024.

Shijie Zhou, Hui Ren, Yijia Weng, Shuwang Zhang, Zhen Wang, Dejia Xu, Zhiwen Fan, Suya You, Zhangyang Wang, Leonidas J. Guibas, and Achuta Kadambi. Feature4x: Bridging any monocular video to 4d agentic AI with versatile gaussian feature fields. In *CVPR*, pp. 14179–14190. IEEE, 2025.

Zrporz. Autoseg-sam2, 2024. URL `https://github.com/zrporz/AutoSeg-SAM2`. Automated image segmentation tool based on Segment Anything Model (SAM).