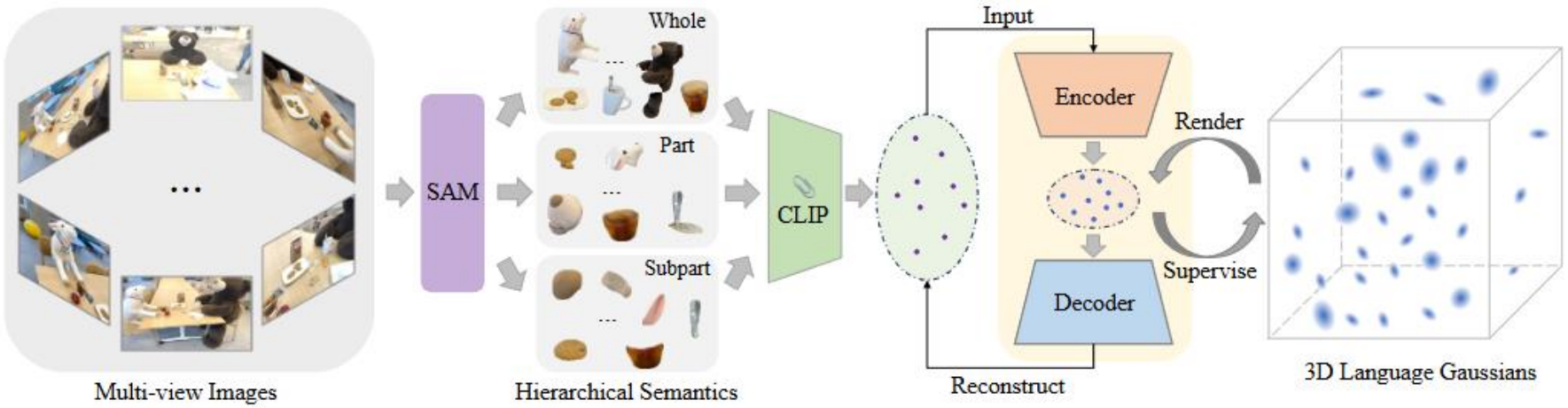
Feature Distillation-based Methods

LangSplat: 3D Language Gaussian Splatting

2024 CVPR

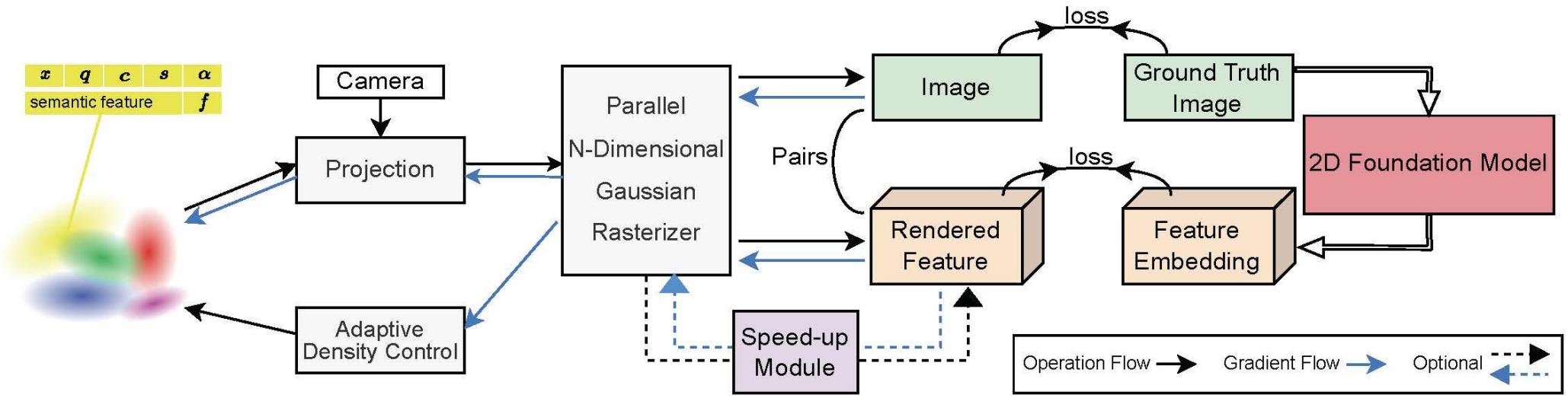
子部分、部分、整体。3D语言场



Feature 3DGS: Supercharging 3D Gaussian Splatting to Enable Distilled Feature Fields

2024 CVPR

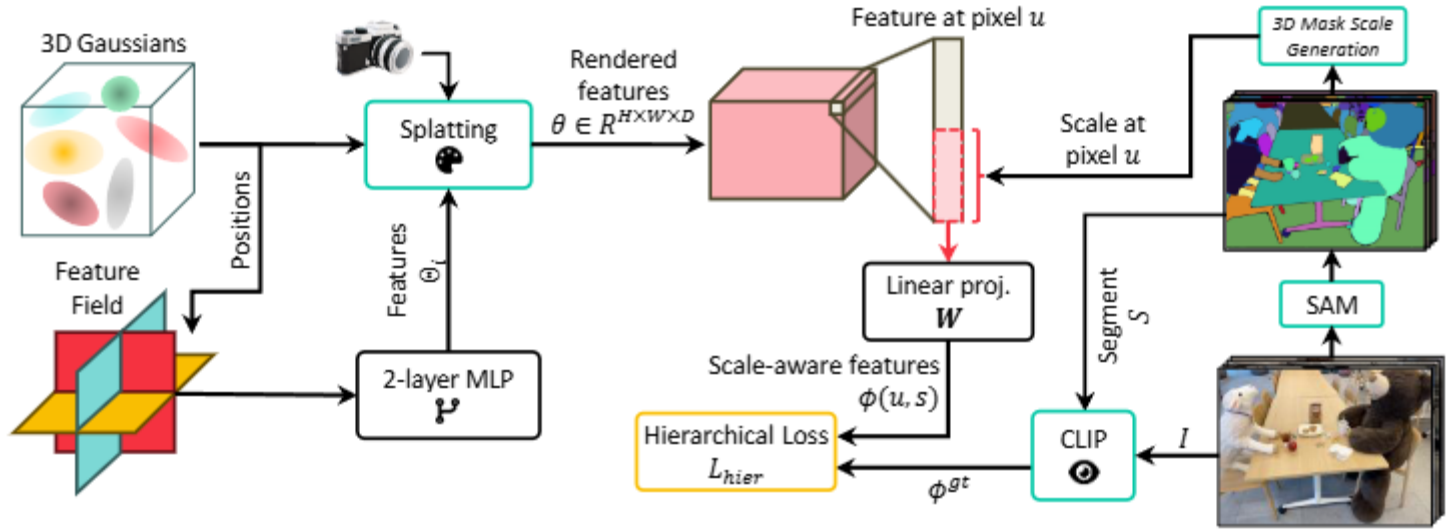
2D蒸馏到3D　LSeg SAM　　维度大



N2F2: Hierarchical Scene Understanding with Nested Neural Feature Fields

2024 ECVA

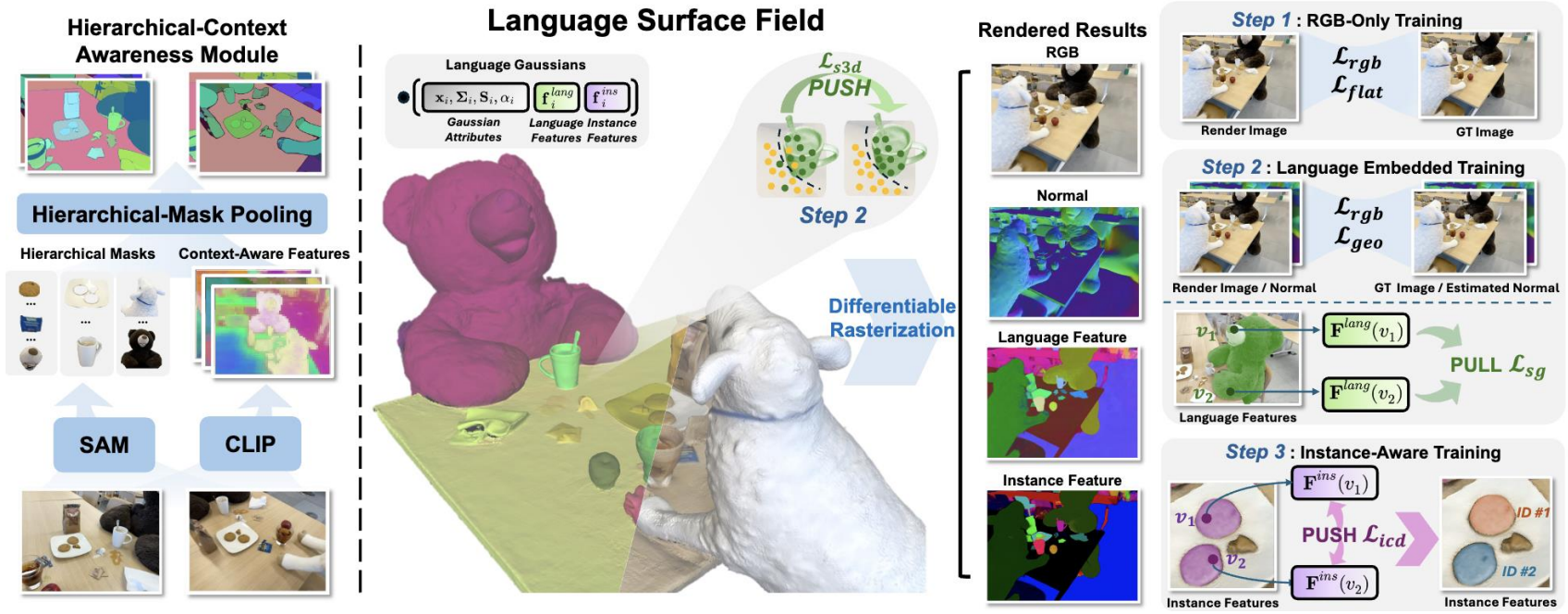
加入尺度概念(前大物体, 后小物体)　$\mathrm{F}_{512} \rightarrow F_{64}$



LangSurf: Language-Embedded Surface Gaussians for 3D Scene Understanding

2024 arXiv

空间感知语义监督 $\mathcal{L}_{s3d} = \sum_{j=1}^{N} \sum_{k=1}^{N_k} f_j^{lang} \cdot \left(-log \frac{f_j^{lang}}{f_k^{lang}} \right)$ 物理位置相邻的两个点，它们的语义（语言特征）也应该是相似的



几何正则化$\mathcal{L}_{geo}$：让高斯点贴在表面上，使用预训练模型估计图像法向量逼迫高斯球变扁、变平，整齐排列在物体表面上

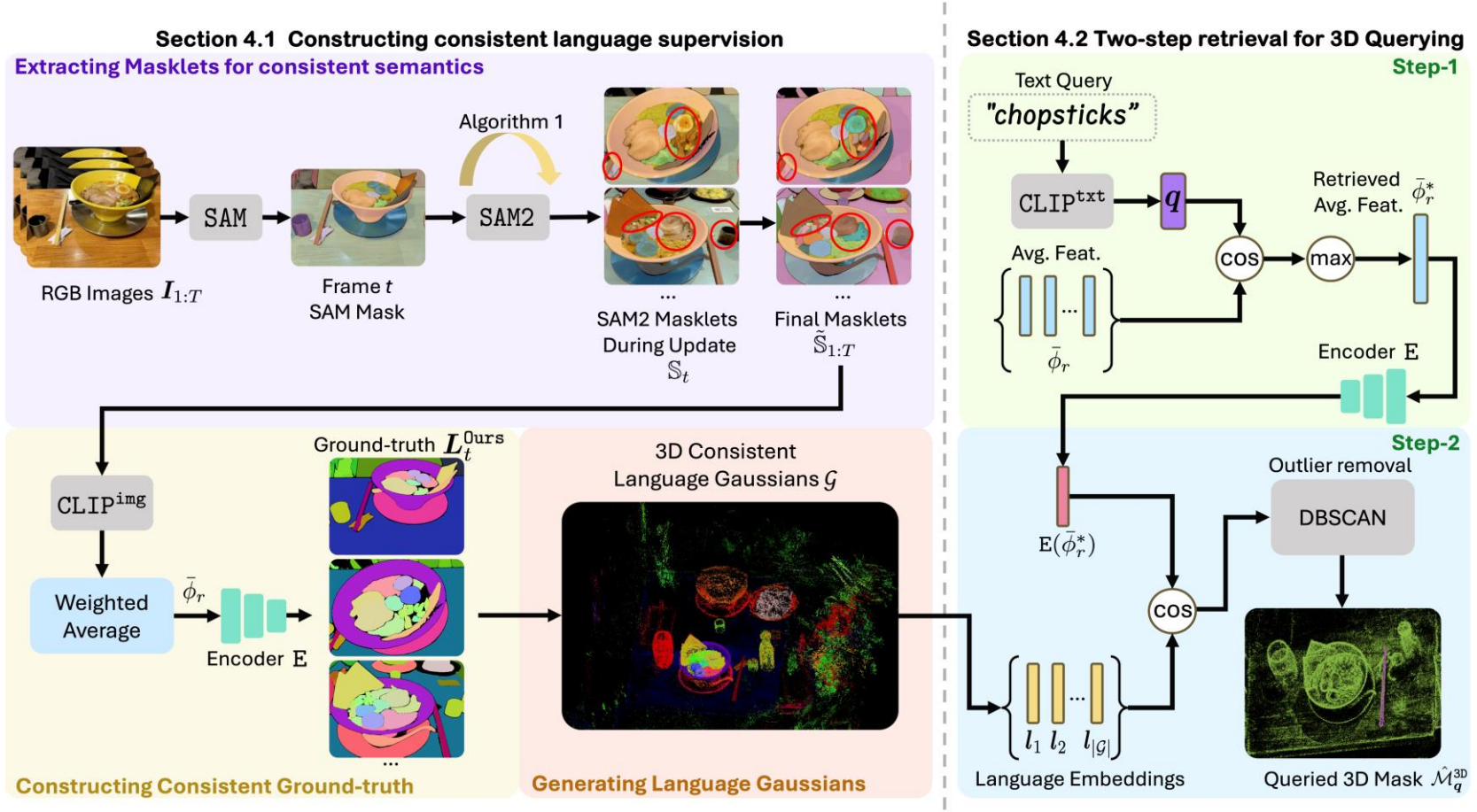
先CLIP再SAM

实例分割实例特征取平均得到一个“指纹” $z_i^{\{ins\}}$

让不同物体的“指纹”尽量不一样 $\mathcal{L}_{icd} = \sum_{j=1}^{M} \sum_{k \neq j} \text{ReLU}\left(D_{min} - \left|\left| z_j^{ins} - z_k^{ins} \right|\right|_2^2 \right)$

Semantic Consistent Language Gaussian Splatting for Point-Level Open-vocabulary

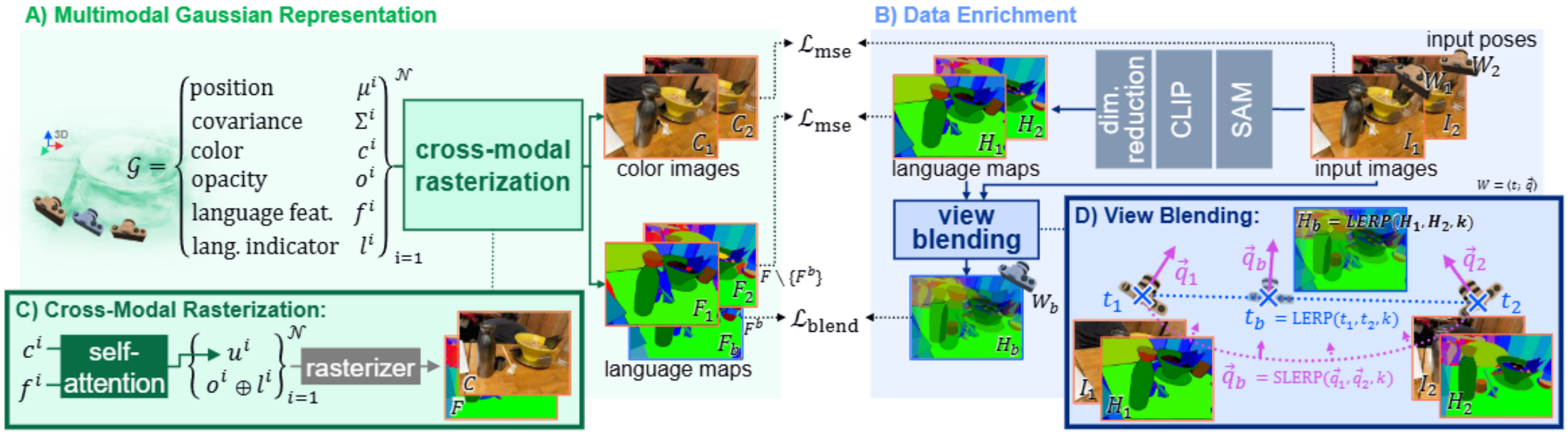
2025 arXiv



3D VISION-LANGUAGE GAUSSIAN SPLATTING

2025 ICLR

X 没有代码

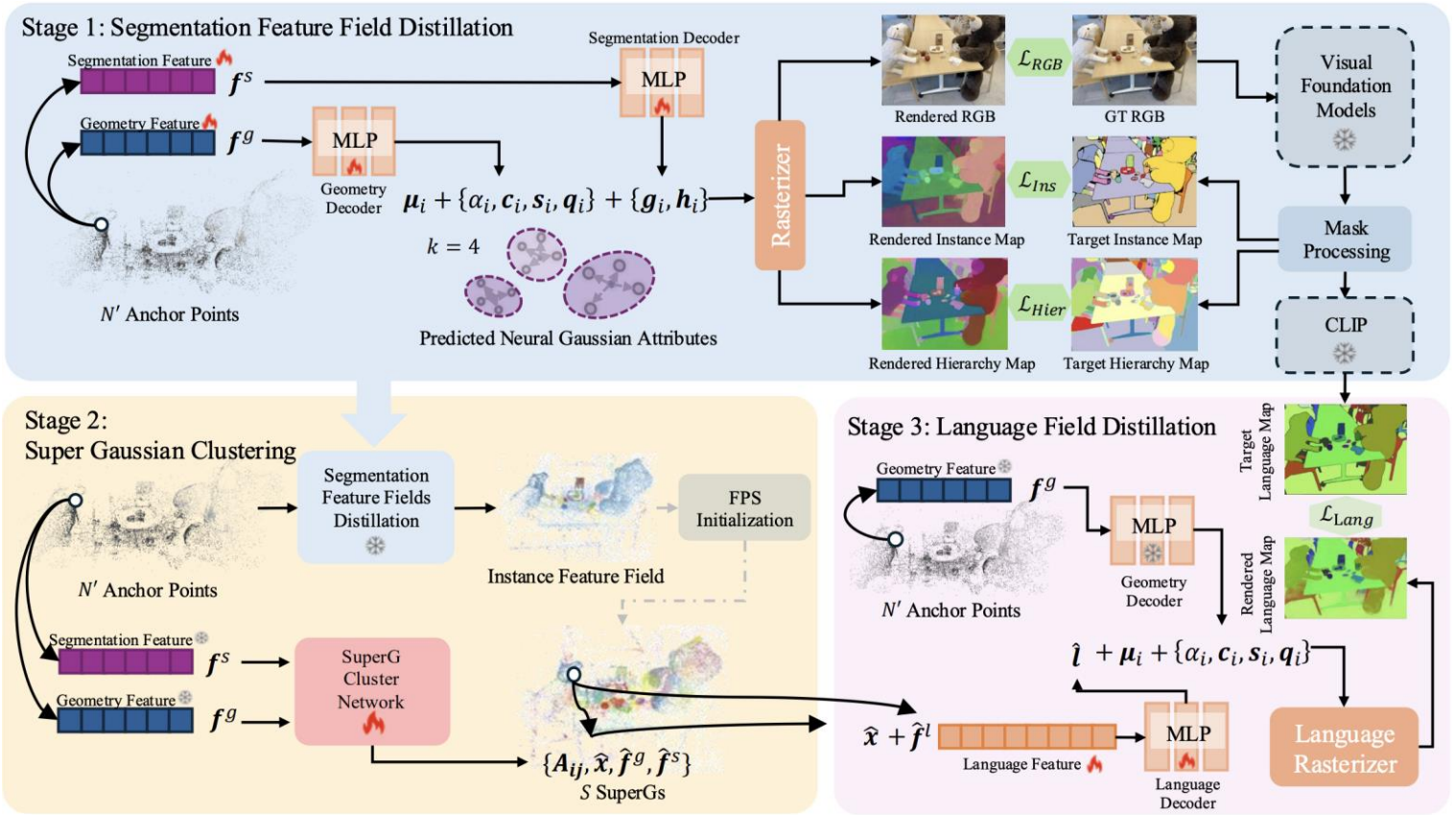


SuperGSeg: Open-Vocabulary 3D Segmentation with Structured Super-Gaussians

2025 3DV

实例特征g 层级特征 h鼓励同一掩码内的特征相似，而与其他掩码的特征相异 $\mathcal{L}_{Ins} = -\frac{1}{|\mathcal{M}|}\sum_{p=1}^{|\mathcal{M}|}\sum_{t=1}^{|\hat{\mathbf{g}}_p|}\log\frac{\exp(\hat{\mathbf{g}}_p^t \cdot \bar{\mathbf{g}}_p/\tau_p)}{\sum_{q=1}^{|\mathcal{M}|}\exp(\hat{\mathbf{g}}_p^t \cdot \bar{\mathbf{g}}_q/\tau_q)}$

L_{Hier}:SAM生成的 2D 掩码通常是重叠的这个损失函数能让模型自动理清这些 Mask 的嵌套关系



SuperG: 聚类锚点
最远点采样(Farthest Point Sampling, FPS)从锚点中初始化S个SuperGs

A_{ij} 第i个锚属于第j个超高斯的概率（或关联强度）$\hat{x}$是坐标

相同的物体在不同位置可能含义不同, 所以加$\hat{x}$

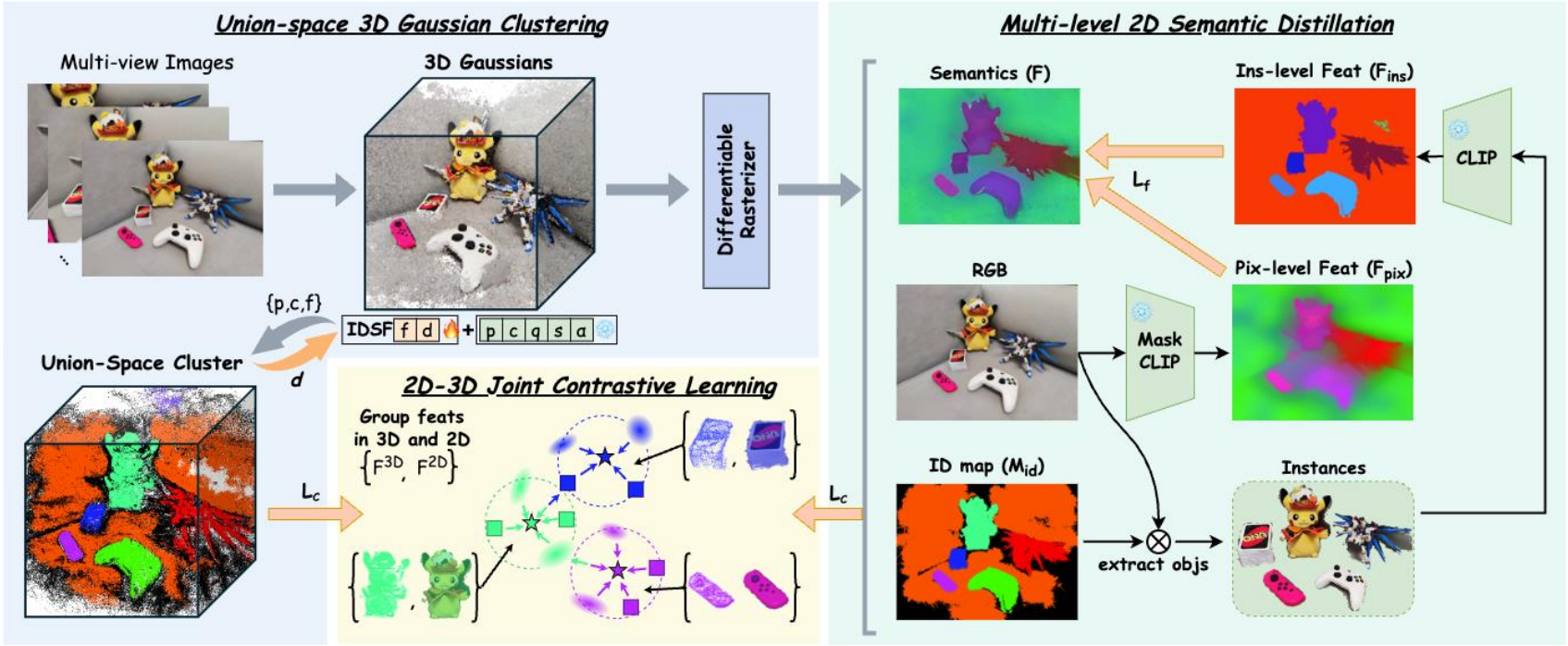
Bootstraping Clustering of Gaussians for View-consistent 3D Scene Understanding (FreeGS)

2025 AAAI

$$L = L_F + \lambda_C L_C + \lambda_S L_S$$

$$\mathcal{L}_F = \|\mathbf{F}_{pix} - \hat{F}\|_1 + \gamma\|\mathbf{F}_{ins} - F\|_1$$

$f_i \in R^D$视图无关的语义向量
$d_i \in R$ 跨视图实例索引



使用MaskCLIP从渲染图像I中提取2D空间视觉-文本对齐特征F_{pix}。
并使用FeatUP将特征分辨率提高到224×224,
将渲染的F通过卷积层降采样到相同分辨率$\hat{F}$

$$\mathcal{L}_S = \frac{1}{TK}\sum_{i=1}^{T}\sum_{j=1}^{K}\left(1 - \mathrm{sim}(f_i, f_j)\right)$$

约束相邻高斯语义来提高聚类组的完整性

将 3D 聚类生成的实例ID渲染为 2D 掩码，并以此为桥梁对齐2D图像与3D高斯的语义特征，无监督对比损失。

$$\mathcal{L}_C^{(i)} = -\sum_{j=1}^{|\mathcal{F}^{\mathcal{G}_i}|}\log\frac{\sum_{\mathcal{G}_{i+}}\exp\left(\mathrm{sim}\left(\mathcal{F}_j^{\mathcal{G}_i}, \overline{\mathcal{F}}^{\mathcal{G}_{i+}}\right)/\tau\right)}{\sum_{\mathcal{G}_{i-}}\exp\left(\mathrm{sim}\left(\mathcal{F}_j^{\mathcal{G}_i}, \overline{\mathcal{F}}^{\mathcal{G}_{i-}}\right)/\tau\right)}$$

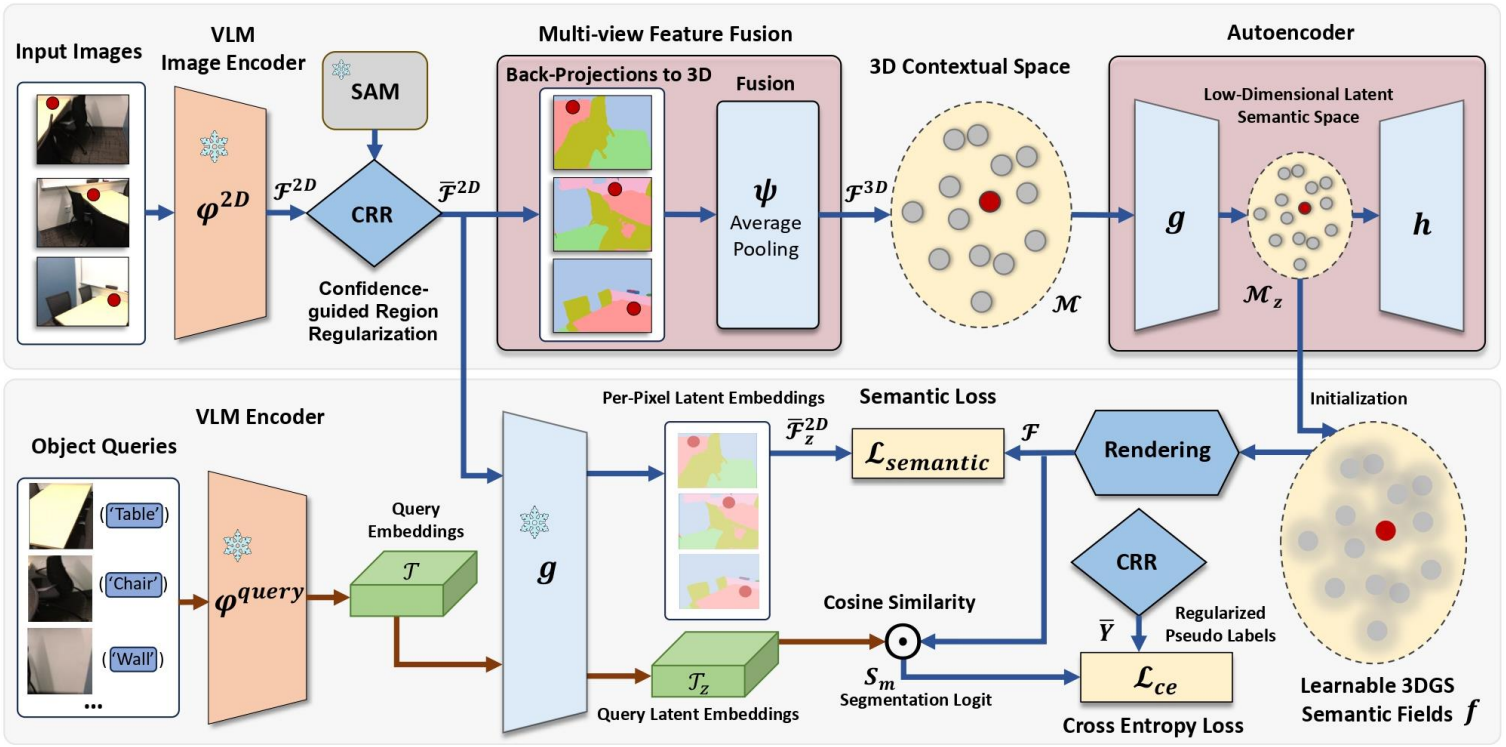
econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians

2025 ICLR

$$L = L_{color} + \lambda_{2d} L_{ce} + \lambda_{sem} L_{semantic}$$

$$\mathcal{L}_{ae} = \mathcal{L}_{l2}(f_p^{3D}, o_p^f) + \mathcal{L}_{ce}(\hat{y}, \cos < o_p^f, T >) + \mathcal{L}_{ce}(\hat{y}, \cos < z_p^f, g(T) >)$$

2D 提纯(置信度非常高的像素) -> 3D 凝聚(深度图Colmap 全部投影到 3D 空间，并进行平均池化) -> 2D 纠偏(多数投票确定每个 2D 区域最稳健的语义标签,画一个边界框（Bounding Box），把这个框作为 Prompt 发给 SAM)



L_{ae}:高低维度编码器损失
L_{l2}:重建损失
o_p^f:高维 z_p^f:低维
768 -> 6

$$L_{semantic} = L_2(\mathcal{F}, \overline{F_z^{2D}}) \qquad L_{ce} = CE(S_m, \bar{Y})$$

ReferSplat: Referring Segmentation in 3D Gaussian Splatting

提出Referring 3D Gaussian Splatting (R3DGS)，并构建了一个新的数据集 RefLERF 来支持 R3DGS 的未来研究

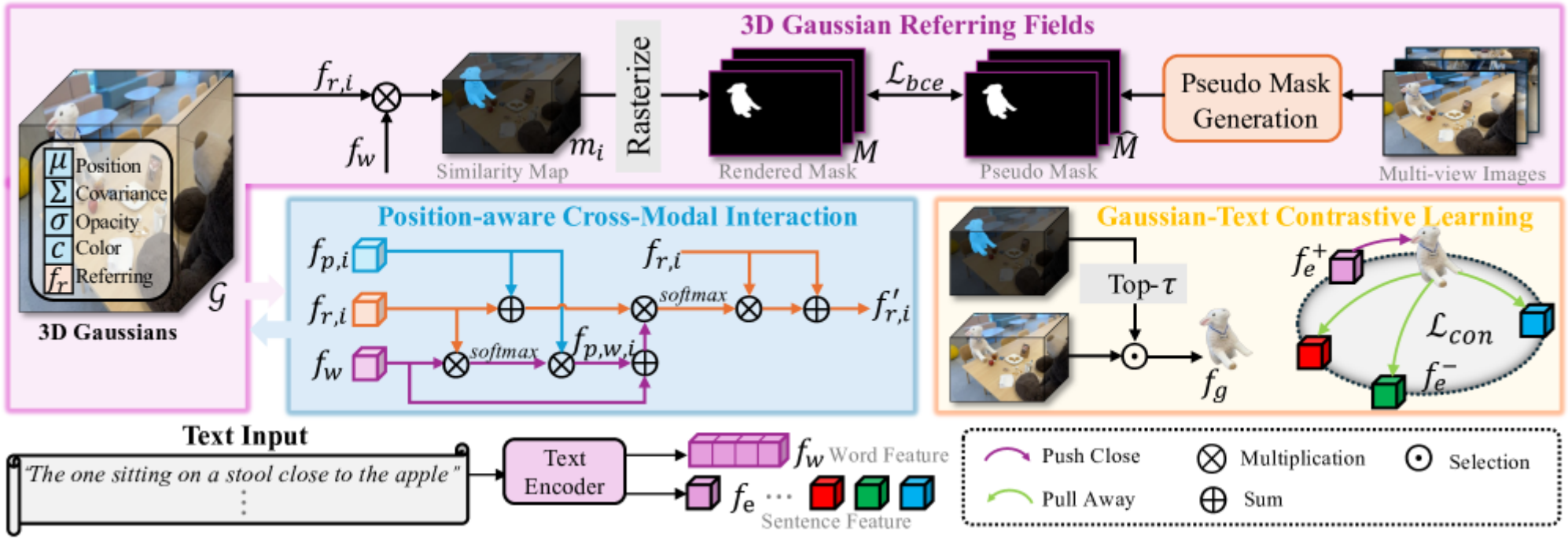
2025 ICML

$$\mathcal{L}_{\text{loss}} = \mathcal{L}_{\text{bce}} + \lambda \mathcal{L}_{\text{con}}$$

$$\mathcal{L}_{\text{bce}} = -\sum_{v} [\hat{y} \log y + (1 - \hat{y}) \log(1 - y)],$$

模型会看很多个候选答案。如果一个答案不仅自己很有信心γ高，而且和其他视角的答案长得很像(IoU 高)，那它就是我们要找的“伪真值” $\hat{y}$

$$p_k = \sum_{j=1}^{K} \text{IoU}(\gamma_k \hat{M}_k, \gamma_j \hat{M}_j)$$



让高斯点的身份证（语义）和它的坐标（位置）深度绑定，学会看方向

$$f'_{r,i} = f_{r,i} + \text{Softmax}\left(\frac{(f_{r,i} + f_{p,i})(f_w + f_{p,w,i})^T}{\sqrt{D}}\right) f_w$$

$$L_{con} = -\frac{1}{|P|} \sum_{f_e^+ \in P} \log \frac{\exp(f_g \cdot f_e^+)}{\sum_{f'_e \in \{P,N\}} \exp(f_g \cdot f'_e)}$$

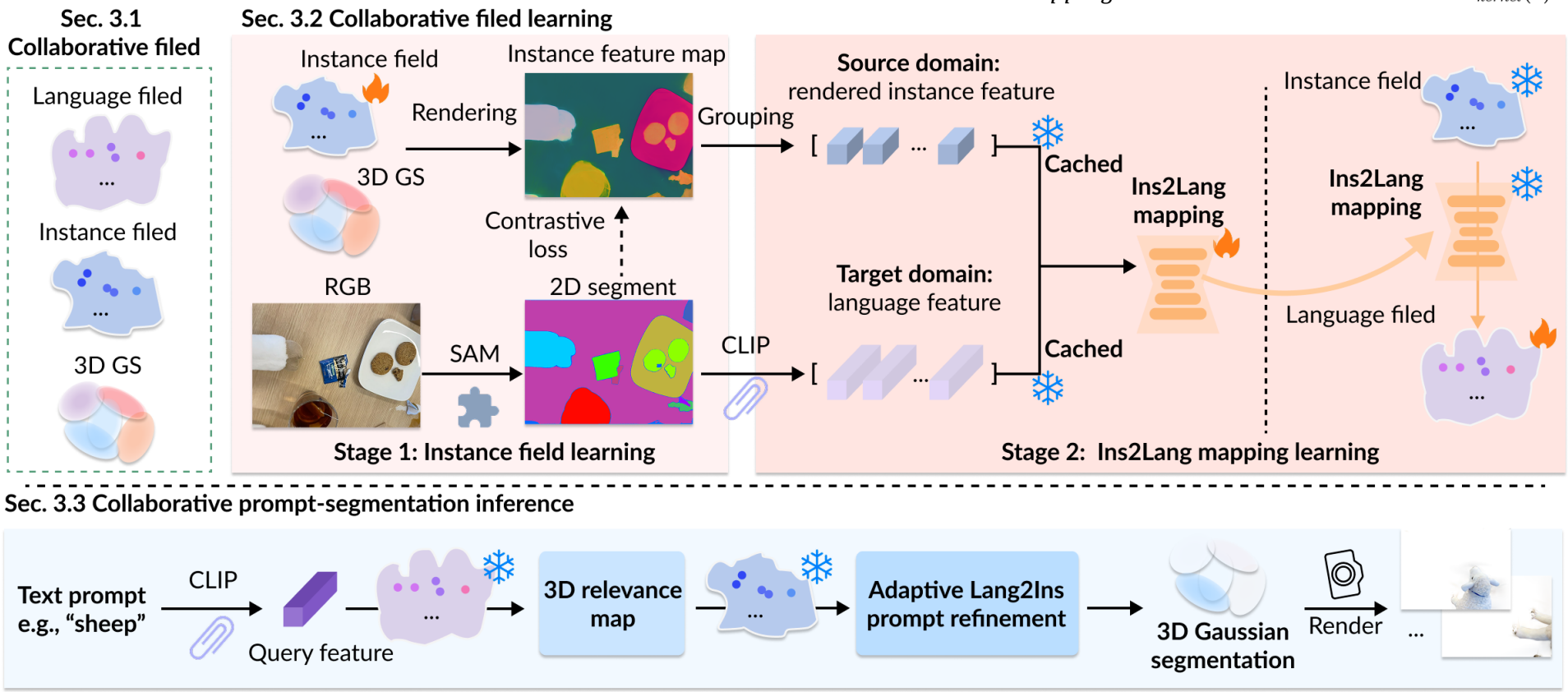
COS3D: Collaborative Open-Vocabulary 3D Segmentation

2025 NeurIPS

InfoNCE 损失函数：I_u是投影迫使同一 Mask 覆盖下的 3D 高斯点学习到了高度相似的特征向量，从而构建了一个“实例场

$$L_{ins} = -\frac{1}{|\Omega|}\sum_{\Omega_j \in \Omega}\sum_{u \in \Omega_j} \log \frac{\exp(\mathrm{sim}(I_u, \bar{I}_j))}{\sum_{\Omega_l \in \Omega} \exp(\mathrm{sim}(I_u, \bar{I}_l))}$$

$$L_{mapping} = |L_m - \Phi_{network}(I_m)| \qquad \Phi_{kernel}(I) = \frac{\sum_{m=1}^{M} \exp\left(-\frac{\|I-I_m\|^2}{2\sigma^2}\right) L_m}{\sum_{m=1}^{M} \exp\left(-\frac{\|I-I_m\|^2}{2\sigma^2}\right)}$$



I_m：输入实例特征 I_m和L_m来自第二阶段准备好的M个训练对（同上文定义）。它们充当了“知识库”

Language Embedded 3D Gaussians for Open-Vocabulary Scene Understanding

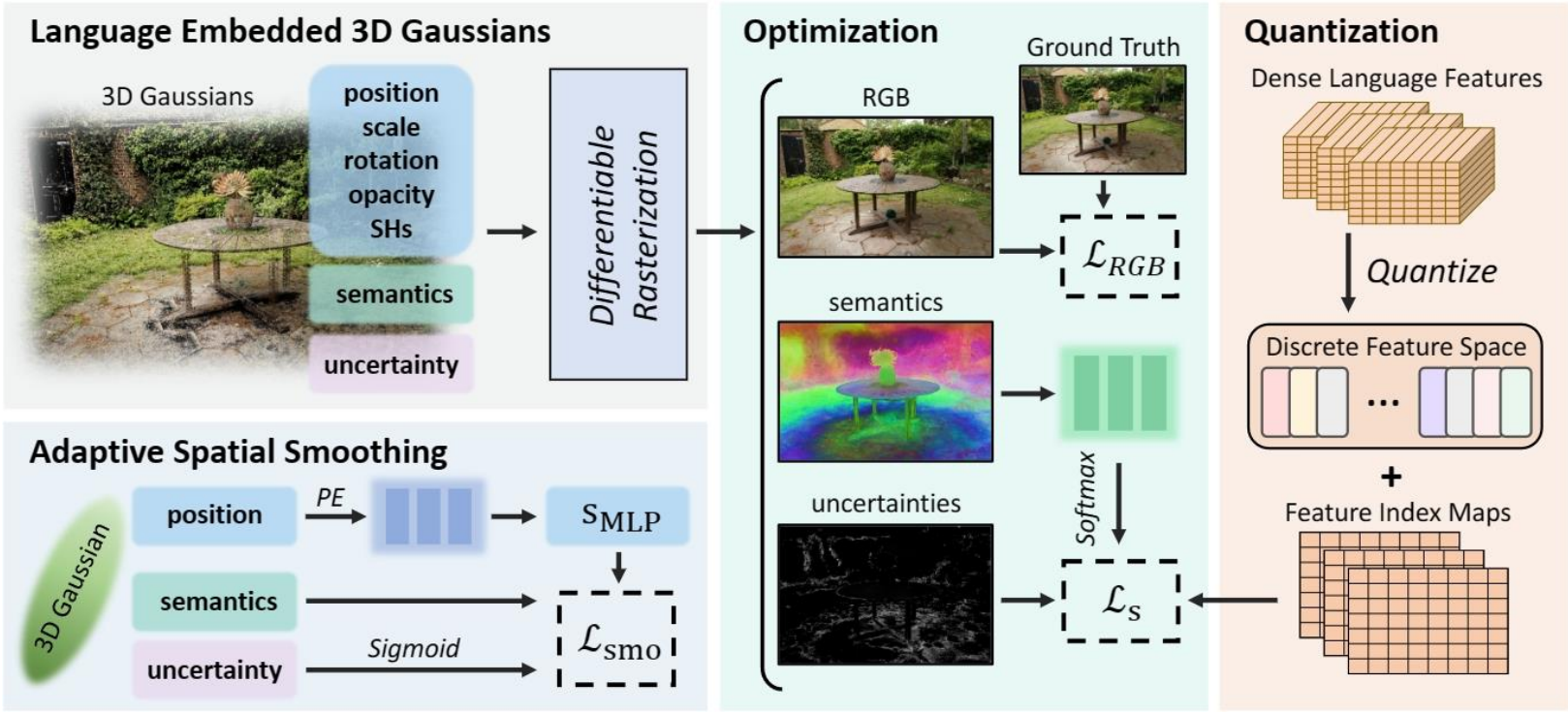
解决“点云表示固有高频偏差”和“多视图不一致性”的关键。作者利用坐标 MLP 的低频感应偏差（Inductive Bias）来规范高斯点的语义不确定是CE损失得到的

2024 CVPR

总损失: $L = \lambda_s L_s + \lambda_{smo} L_{smo}$

$$\mathcal{L}_{\text{smo}} = \|s_{\text{MLP}} - s_G^*\|_2 + \max(u_G^*, w_s)\|s_{\text{MLP}}^* - s_G\|_2$$

$$L_{cos}(F_i) = \left(1 - \cos\langle F_{CLIPi} \cdot \widehat{F_{CLIPi}}\rangle\right) + \lambda_{DINO}\left(1 - \cos\langle F_{DINOi} \cdot \widehat{F_{DINOi}}\rangle\right)$$



$$L_q = \lambda_{cos} L_{cos} + \lambda_{lb} L_{lb}$$

$$L_{lb} = \sum_{i=1}^{N} (r \circ p)$$

借鉴了 Switch Transformer 的思想，防止“量化坍缩”（即所有特征都映射到少数几个基向量上）

$$L_s = \lambda_{CE} L_{CE} + \lambda_u L_u \quad L_{CE} = \frac{\sum CE(\widehat{M}, M) \circ (1 - R_u(G; p_{cam}))}{H \times W} \quad L_u = \frac{\sum R_u(G; p_{cam})}{H \times W}$$

R_u是渲染出的不确定性图。如果某个区域在多视图下表现出强烈的语义不一致（例如反光、遮挡）R_u会变高 <- 不确定高斯内在

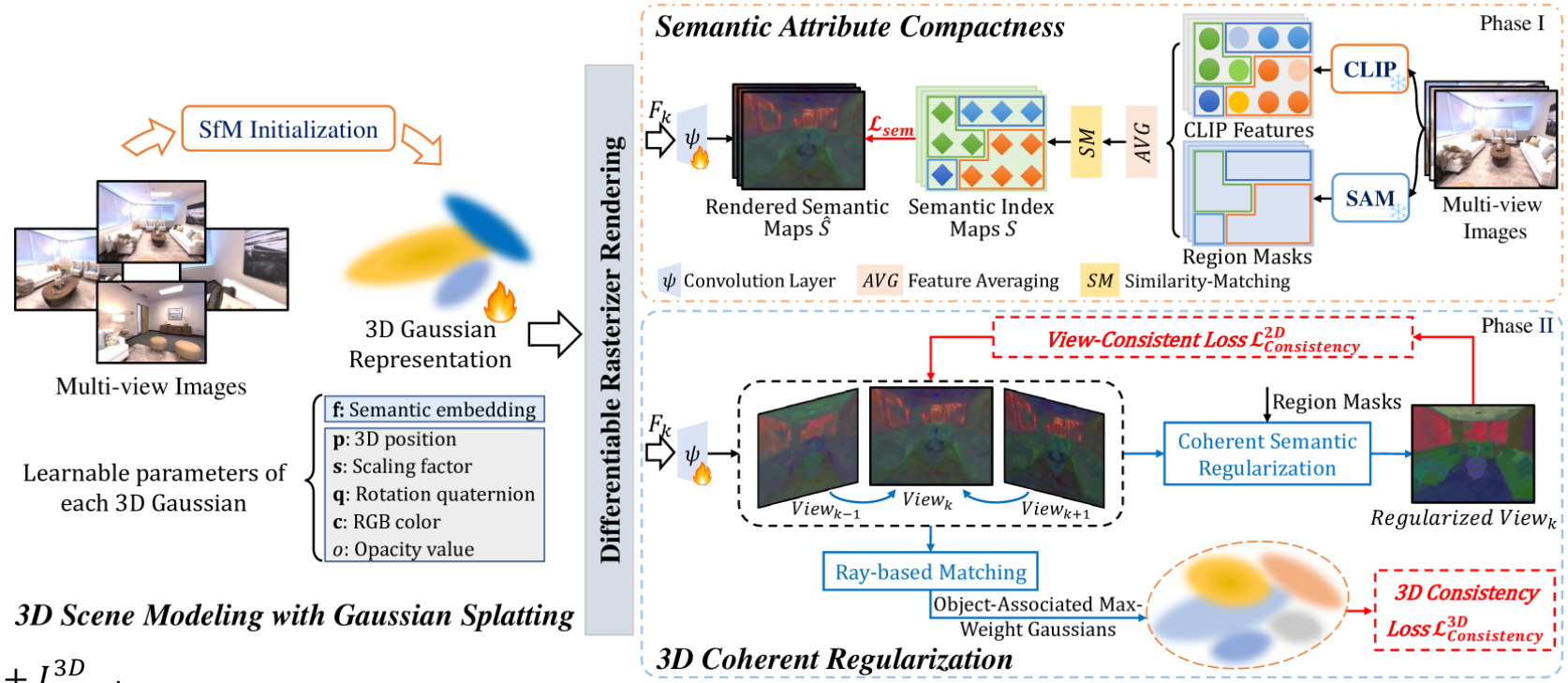
CLIP-GS: CLIP-Informed Gaussian Splatting for View-Consistent 3D Indoor Semantic Understanding

2025 ACM TOMM

如果只用 L_{sem}，模型会陷入 2D CLIP 固有的偏差中

$L_{sem} = L_{ce}(\hat{S}, S)$ 语义损失。当模型已经有了一定的语义基础后，作者不再盲目迷信原始的 2D CLIP 标签（因为它们在多视角下是不一致的），而是引入了自监督的连贯性约束

语义索引3维



$$L_{consistency} = L^{2D}_{consistency} + L^{3D}_{consistency}$$

2D 视图一致性约束(零样本视觉跟踪器) $L^{2D}_{consistency} = L_{ce}(\hat{S}, Z)$

3D 高斯特征一致性约束，对于同一个物体(通过射线匹配找到)它们对应的 3D 高斯基元的语义特征f_i应该非常接近。 $L^{3D}_{consistency} = \sum_{i=1}^{n} M \log \frac{M}{f_i}$

FMGS: Foundation Model Embedded 3D Gaussian Splatting for Holistic 3D Scene Understanding

筛选标准：（下面这些才会有语义特征）
高不透明度（High Opacity）：那些快看不见的透明点就不管了。
足够大（Radius > 2 pixels）：太小的点对语义贡献不大，也排除了。

2024 IJCV

索引使用的是3D坐标 X_i指的是第i个高斯点在 3D 世界空间中的中心位置的坐标，多视角仍然保持一致性

总损失：$L_{total} = \lambda L_{CLIP} + (1-\lambda) L_{DINO} + \gamma L_{pixel}$

在差异小时为 L_2，差异大时为 L_1

$$L_{CLIP} = \begin{cases} 0.5|\hat{F} - F|^2, & \text{if } |\hat{F} - F| < \delta \\ \delta \cdot (|\hat{F} - F| - 0.5 \cdot \delta), & \text{otherwise} \end{cases}$$

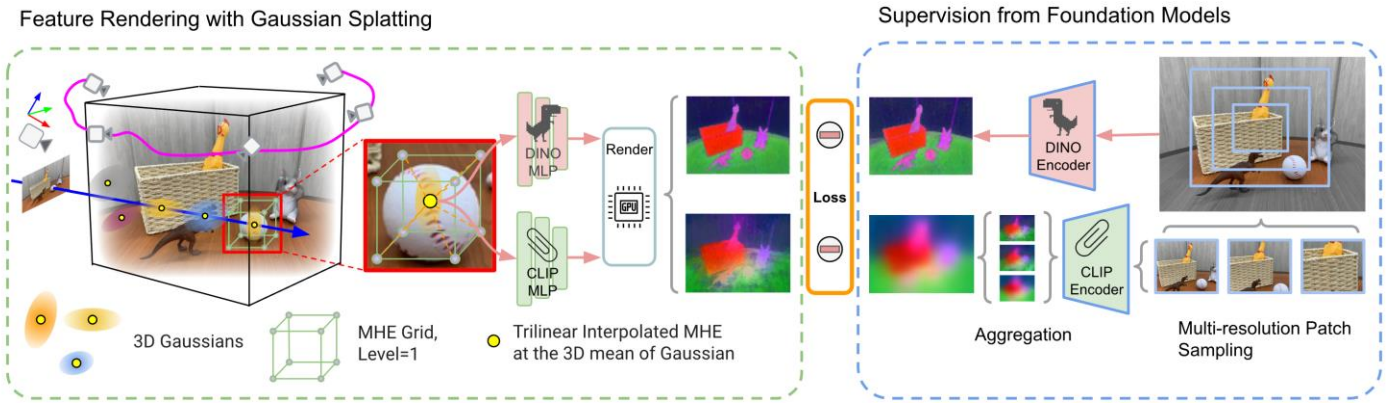
$$\mathrm{L_{DINO}} = \left|\hat{\mathrm{D}} - \mathrm{D}\right|^2$$

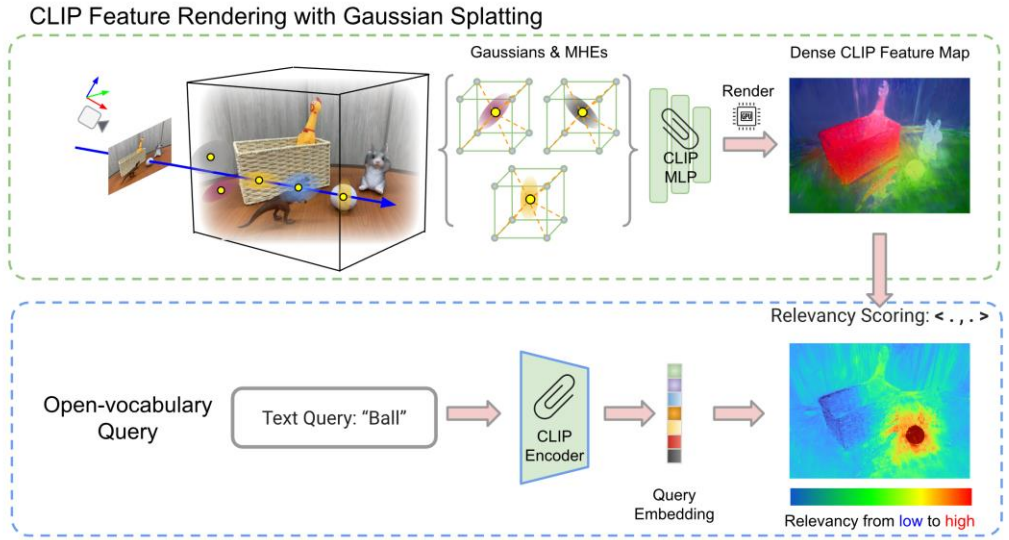
该损失函数强制要求渲染出的CLIP特征之间的点积相似度模式与渲染出的DINO特征模式保持一致，在计算此损失时，会停止DINO特征的梯度回传，从而让CLIP特征去“跟随”DINO展现出的几何边界。

$$L_{pixel} = \frac{1}{K^2 - 1} \sum_{i \in P} \sum_{j \in N(i), j \neq i} |\hat{d}_i^T \hat{d}_j - \hat{f}_i^T \hat{f}_j|$$

CLIP 和 DINO 共享同一个哈希编码器

不同尺寸的裁剪





GOI: Find 3D Gaussians of Interest with an Optimizable Open-vocabulary Semantic-space Hyperplane

2024 ACM MM

$$\mathcal{L} = \mathcal{L}_{\mathcal{T}} + \lambda_{\text{joint}}\mathcal{L}_{joint} + \lambda_{\text{e2e}}\mathcal{L}_{e2e}$$

$$\mathcal{L}_{OSH} = -\frac{1}{P}\sum_{i=1}^{P}[w \cdot \hat{m}_i \log(\sigma(m_i)) + (1-\hat{m}_i)\log(1-\sigma(m_i))]$$

解决特征冗余和多视角不一致性：

$$\mathcal{L}_{\mathcal{T}} = \lambda_{\text{ent}}\mathcal{L}_{ent} + \lambda_{\max}\mathcal{L}_{max}$$

$$\mathcal{L}_{ent} = -\sum_{i=1}^{N} p_i \log(p_i)$$

强制使像素特征只倾向于匹配码本中的某一个条目。这起到了聚类的作用，能够有效去除 2D 特征图中的噪声，使语义边界更加尖锐且一致

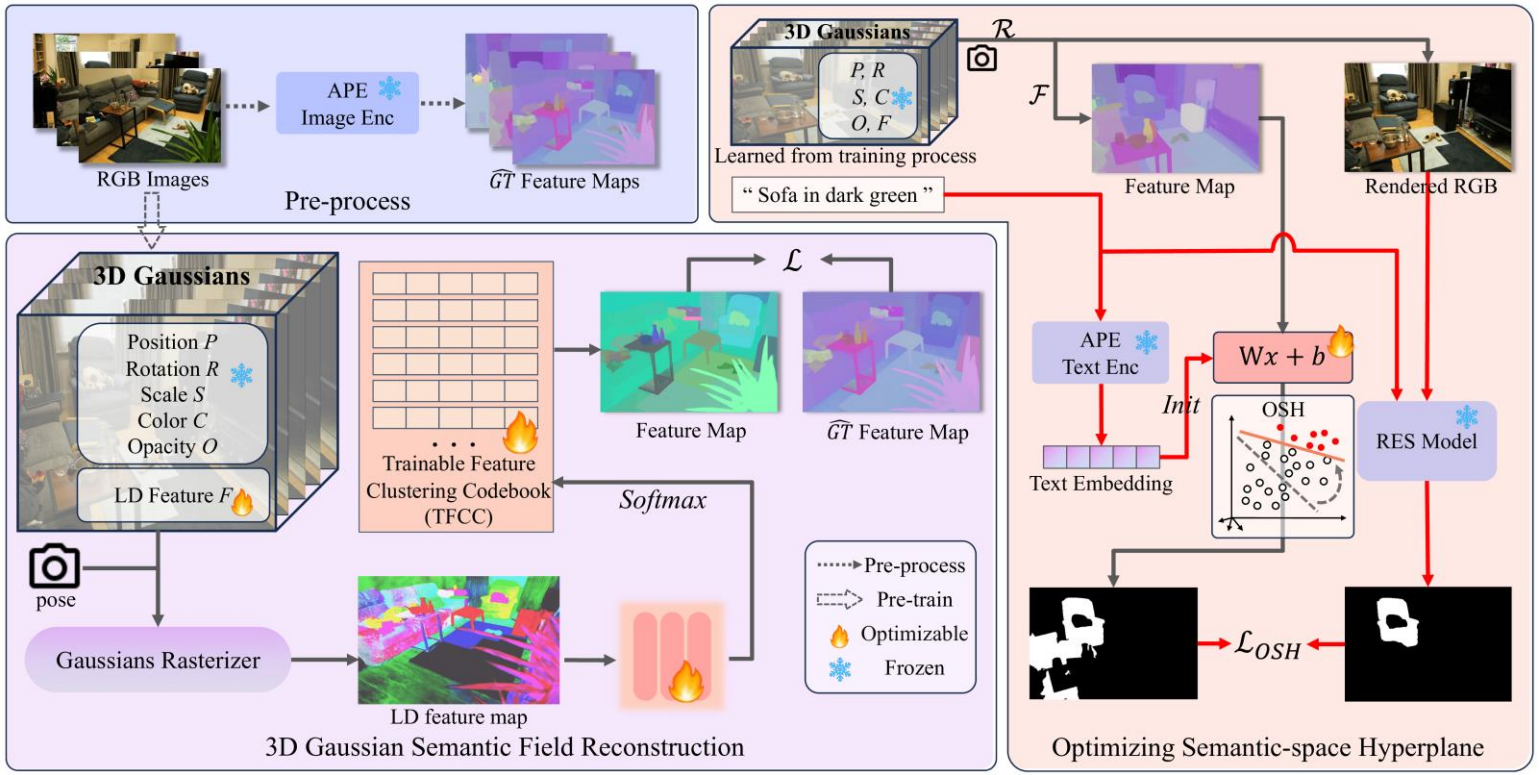
$$p_i = Softmax(\cos(v_{gt}, T[i]) \cdot \tau)$$

v_{gt}是APE提取的真值语义特征；$T[i]$是码本中的第i个条目；τ是退火温度

$$\mathcal{L}_{max} = 1 - \cos(v_{gt}, T[d])$$

$$d = argmax_i(\cos(v_{gt}, T[i]))$$

拉近真值特征与码本中最接近条目的距离，加速码本的收敛，确保码本能准确代表场景中的语义分布



P样本（像素）总数

$\hat{m}_i$由 RES模型(如 Grounded-SAM)生成的 2D 二值掩码，作为伪标签

m_i当前超平面$Wx + b = 0$预测的掩码分数

$\mathcal{L}_{joint} = |e - \text{onehot}(d)|_2^2$ e是低维语义特征通过 MLP 解码器 D后得到的 Logits一个强约束，旨在建立起低维表示与高维语义之间的可靠连接

$\mathcal{L}_{e2e} = 1 - \cos(v_{gt}, v)$直接在最终恢复的语义向量和真值之间进行对比，保证了重建过程的鲁棒性

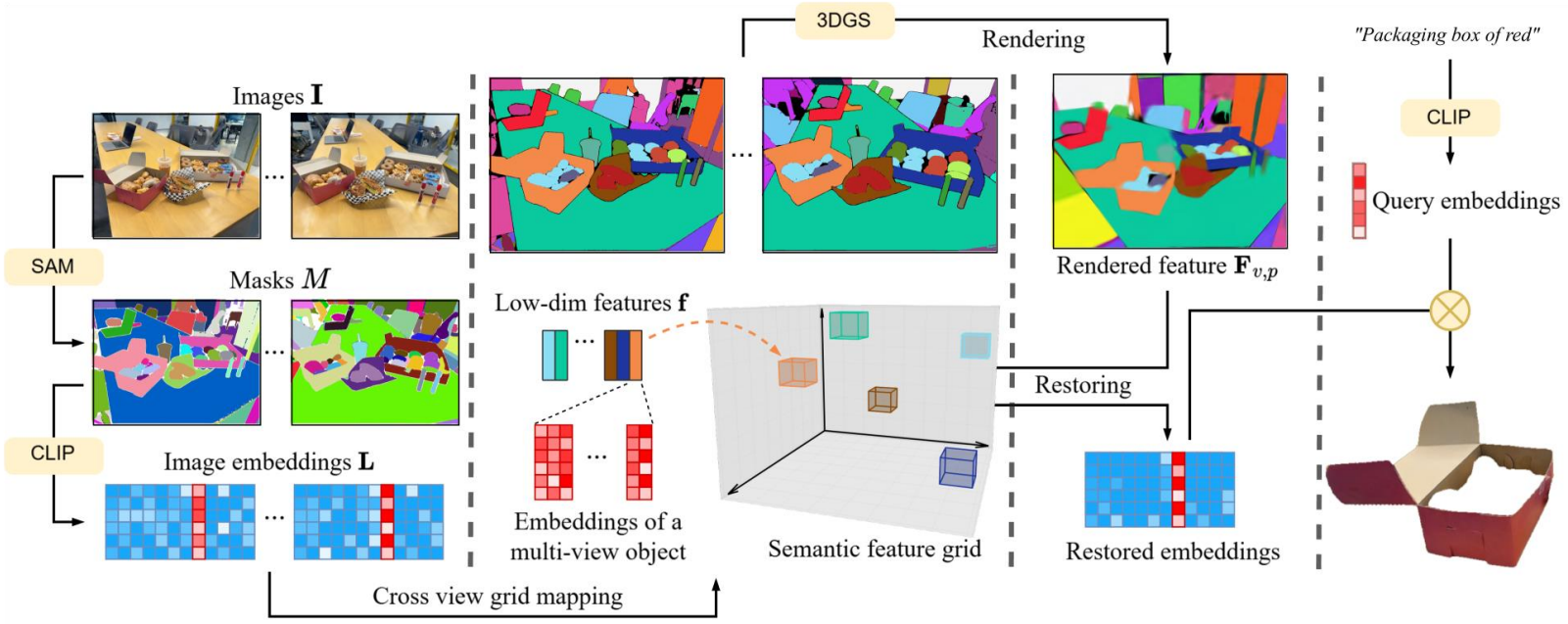
FastLGS: Speeding Up Language Embedded Gaussians with Feature Grid Mapping

2025 AAAI

$$L_f = (1 - \lambda)L_1 + \lambda L_{D-SSIM}$$

跨视角网格映射
使用SIFT和k-近邻（KNN）算法获取掩码间的关键点对应。如果掩码对之间的对应关键点数量超过预设阈值 τ，则它们被视为同一对象的不同视角实例，并分配相同的低维特征

对于关键点较少或不明显的分割掩码，通过计算混合特征（CLIP嵌入L和颜色分布 C）的相似度来匹配。

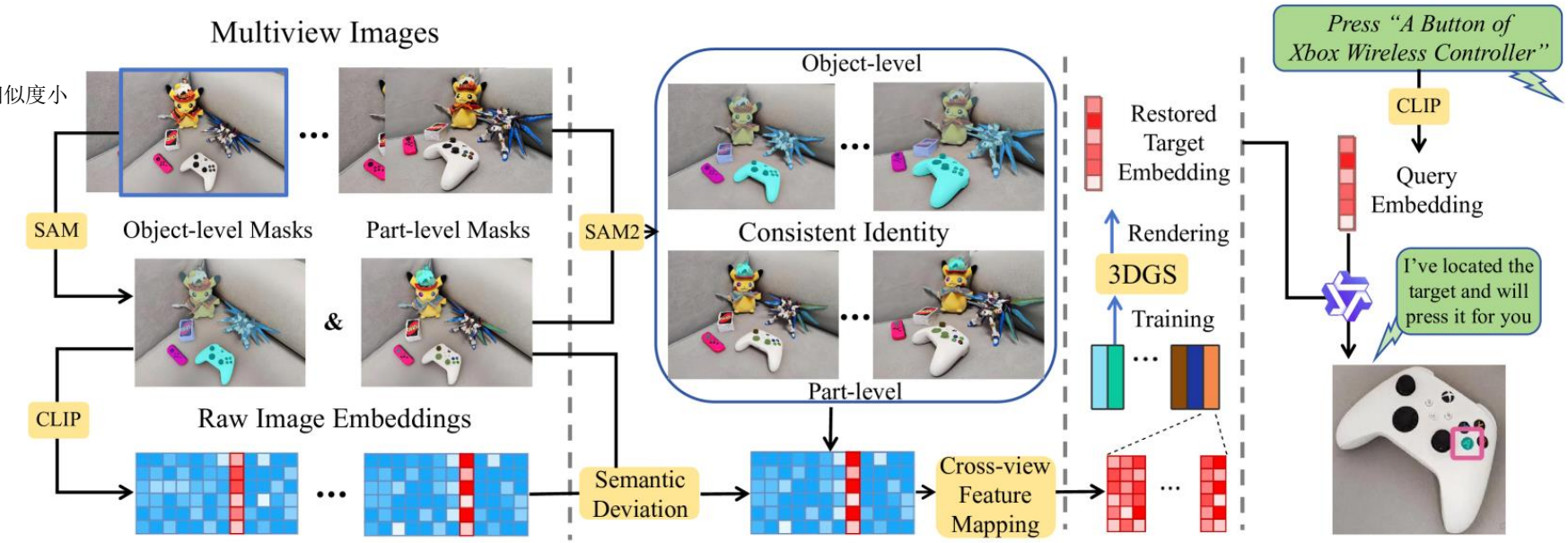


FMLGS: Fast Multilevel Language Embedded Gaussians for Part-level Interactive Agents

2025 arXiv

$$S_{relev} = \min_i \frac{\exp(\phi_{img} \cdot \phi_{query})}{\exp(\phi_{img} \cdot \phi^i_{canon}) + \exp(\phi_{img} \cdot \phi_{query})}$$

ϕ^i_{canoni} 背景提示词 与所有背景相似度小



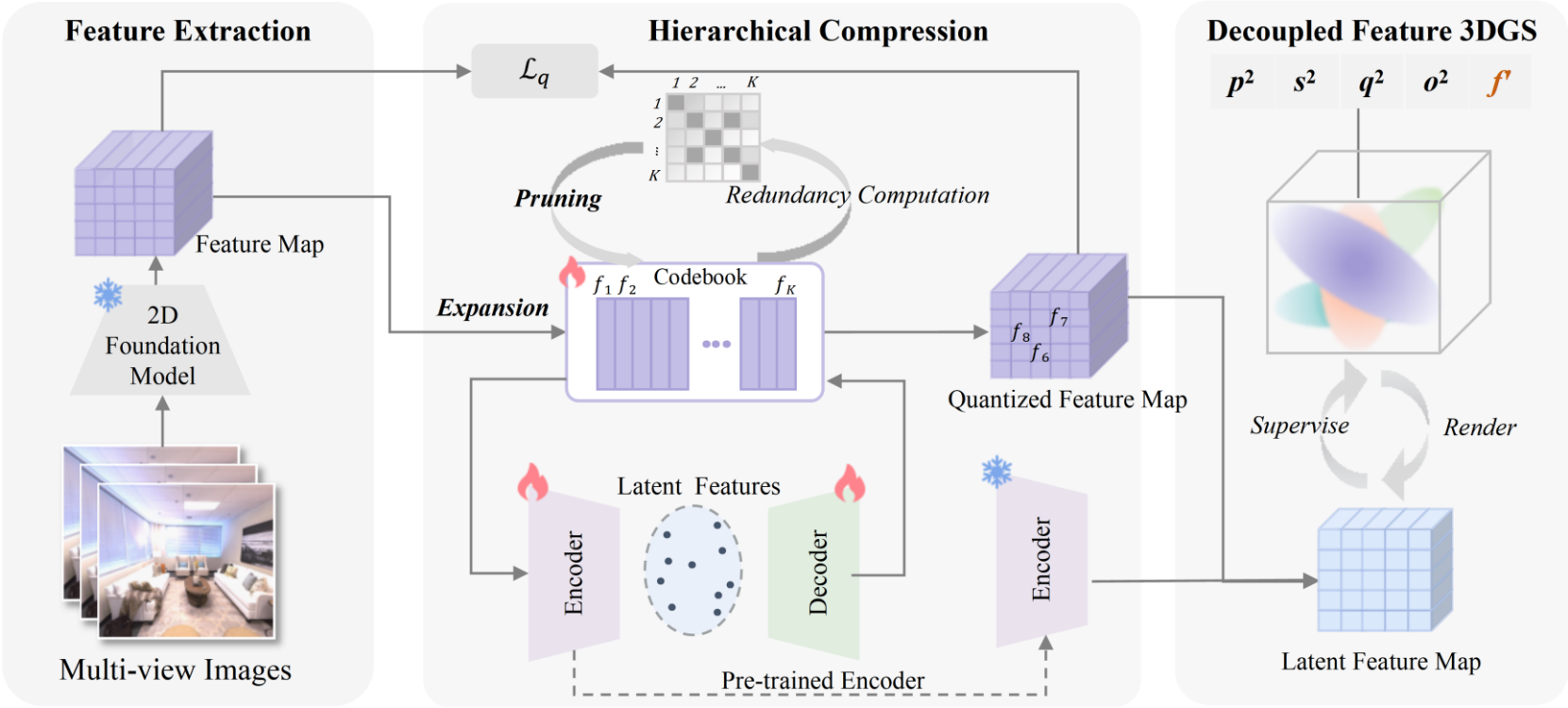
将物体全局特征与部件局部特征进行加权融合 $F'_P = (1-w)F_O + wF_P$，解决了 CLIP “词袋效应”导致的部件归属感缺失和歧义问题 。

Efficient Decoupled Feature 3D Gaussian Splatting via Hierarchical Compression

$L_{ae} = \left|\Psi\left(E(f_e)\right) - f_e\right|^2$
典型的重构损失 它要求原始码词f_e 经过编码器 E 变成 9 维、再经过解码器Ψ还原后，必须和原来的 512 维特征越像越好

2025 CVPR

码本 量化 9维度语义特征



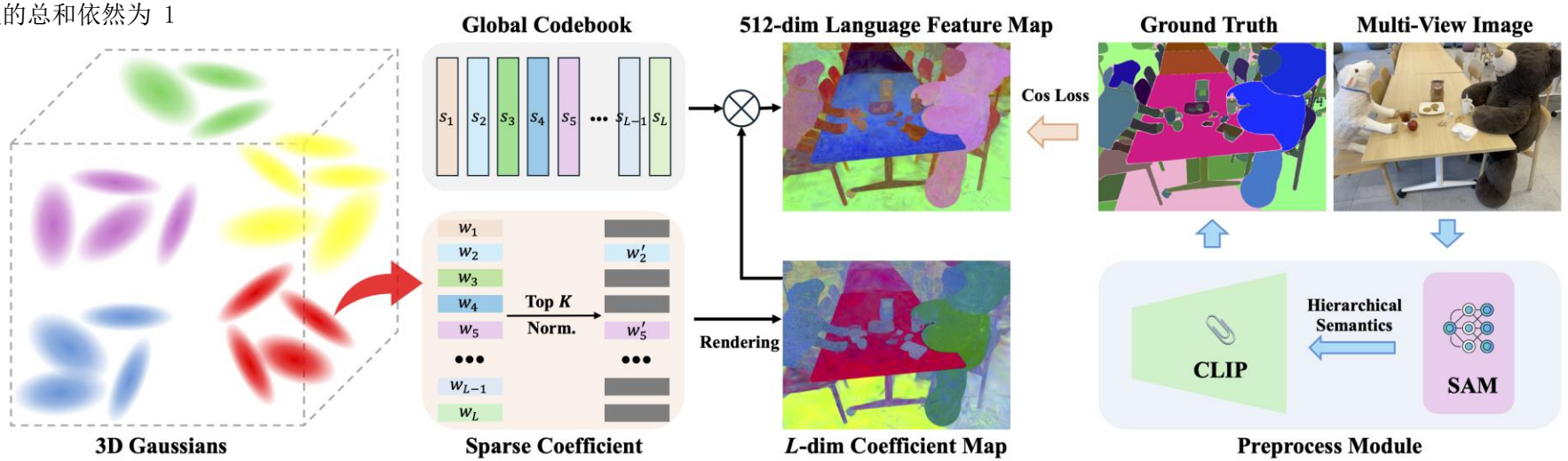
$\mathcal{L}_q = \frac{1}{T}\sum_{\tau=1}^{T}\frac{1}{N_\tau}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\left(1-\cos\left(F_{i,j},\widehat{F_{i,j}}\right)\right)\times\delta\left(M_{i,j},\tau\right)$ 码本生成过程，不能直接平均(小物体会没有)，按类别取平均

LangSplatV2: High-dimensional 3D Language Gaussian Splatting with 450+ FPS

2025 NeurIPS

在训练和推理时，系统并不会让一个点同时对应 64 个向量 它只选出最相关的 4 个
Softmax 归一化：先对这 64 个维度的参数做一次 Softmax
保留 Top-K：只保留数值最大的前 4 个权重，剩下的 60 个全部强制设为 0
重新归一化：确保这 4 个权重的总和依然为 1

码本 64个512维度



RGB 几何重建损失 $\mathcal{L}_{\mathcal{RGB}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\mathcal{SSIM}}$

多层级余弦相似度损失 $\mathcal{L}_{sem} = \sum_{scale\in\{1,2,3\}}\left(1 - \frac{F_{pred}^{scale}\cdot F_{GT}^{scale}}{\left|F_{pred}^{scale}\right|\left|F_{GT}^{scale}\right|}\right)$

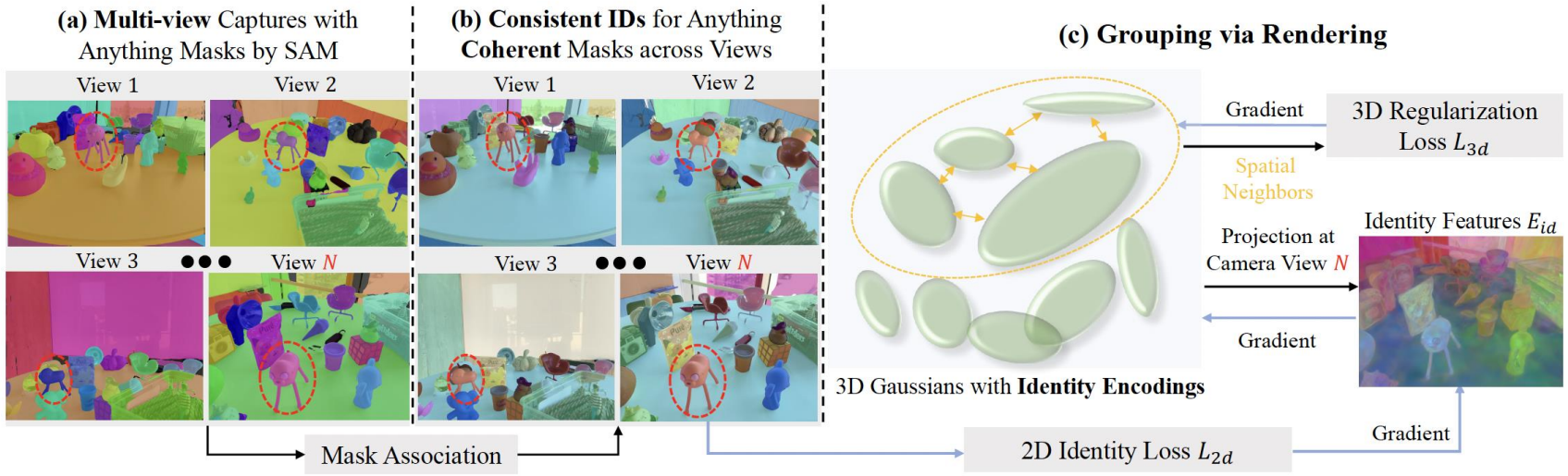
由渲染出的系数图 W 与全局码本 S 相乘得到: $F_{pred} = W \cdot S$

2D Mask Lifting-based Methods

Gaussian Grouping: Segment and Edit Anything in 3D Scenes

2024 ECCV

DEVA KNN相似 16维度身份编码



$$L_{render} = L_{rec} + \lambda_{2d} L_{2d} + \lambda_{3d} L_{3d}$$

L_{2d} 将渲染出的 $E_{\{id\}}$通过一个线性层映射回 K 维(为场景中实例的总数)然后使用标准交叉熵损失进行监督

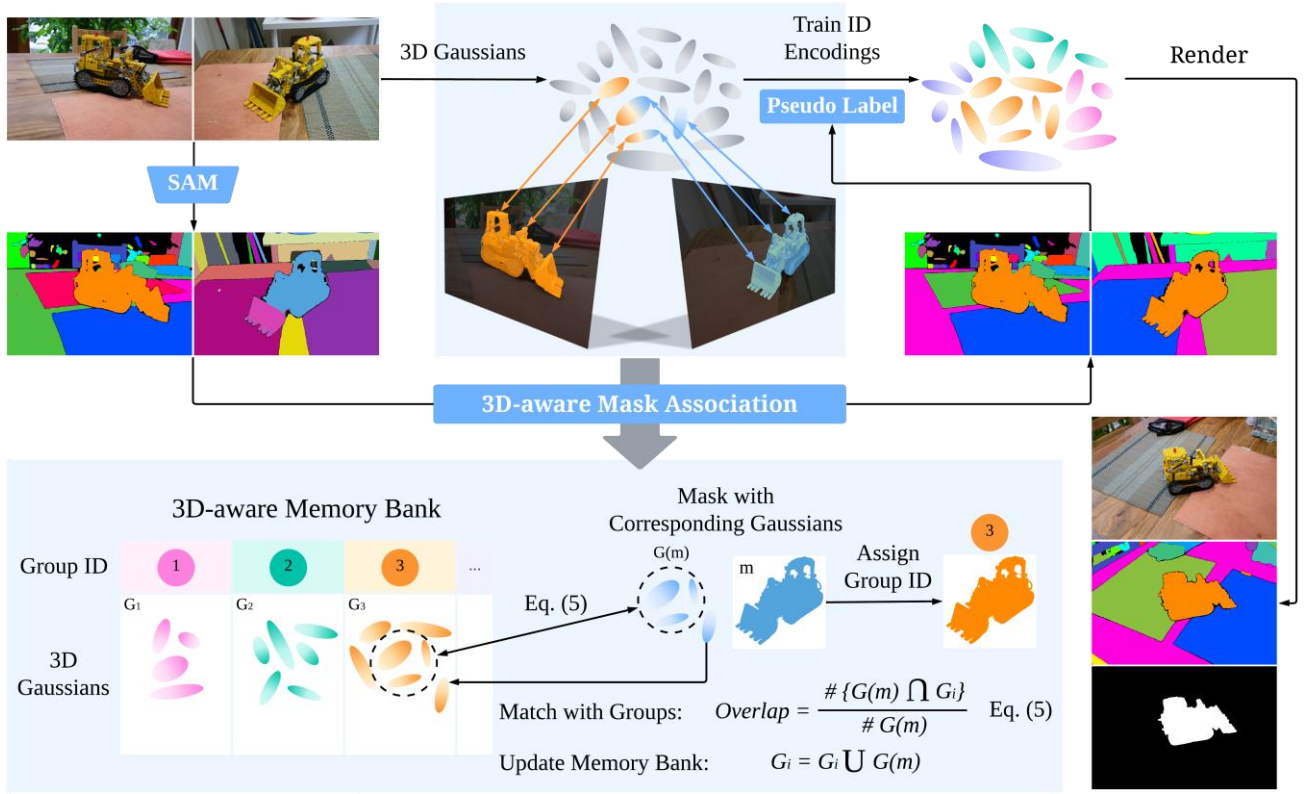
图像重建损失 $L_{rec} = (1-\tau)L_1 + \tau L_{D-SSIM}$

3D 空间一致性正则化损失 与空间最近保持一致：$L_{3d} = \frac{1}{mk}\sum_{j=1}^{m}\sum_{i=1}^{k} D_{kl}(P||Q)$

Gaga: Group Any Gaussians via 3D-aware Memory Bank

2025 arXiv

码本 64个512维度

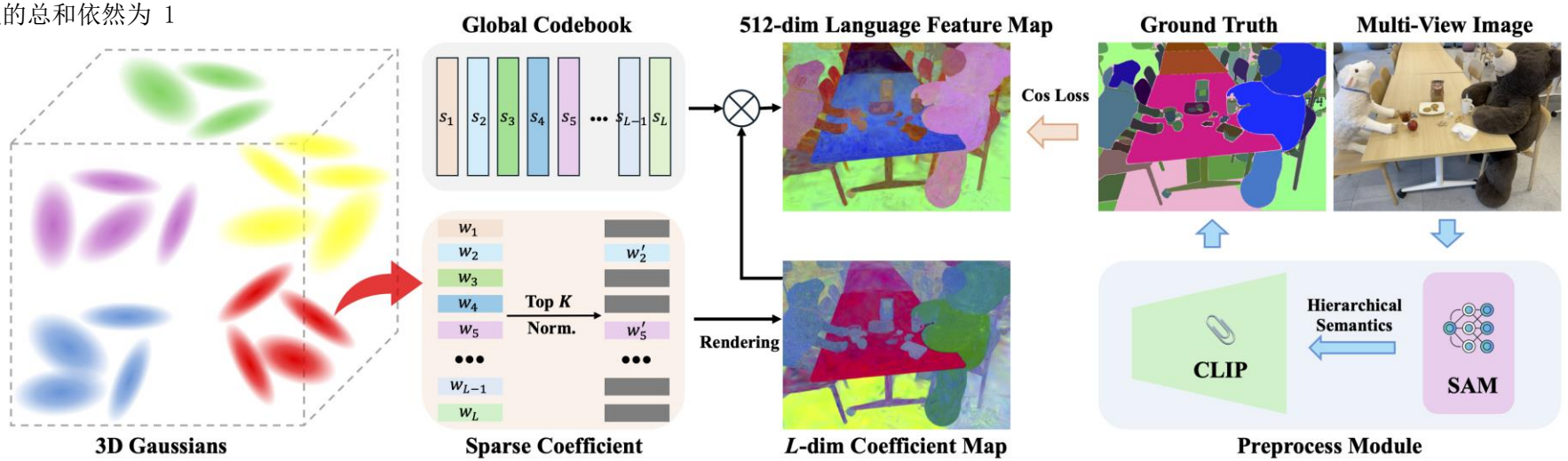


LangSplatV2: High-dimensional 3D Language Gaussian Splatting with 450+ FPS

2025 NeurIPS

在训练和推理时，系统并不会让一个点同时对应 64 个向量 它只选出最相关的 4 个
Softmax 归一化：先对这 64 个维度的参数做一次 Softmax
保留 Top-K：只保留数值最大的前 4 个权重，剩下的 60 个全部强制设为 0
重新归一化：确保这 4 个权重的总和依然为 1

码本 64个512维度



RGB 几何重建损失 $\mathcal{L}_{\mathcal{RGB}} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\mathcal{SSIM}}$

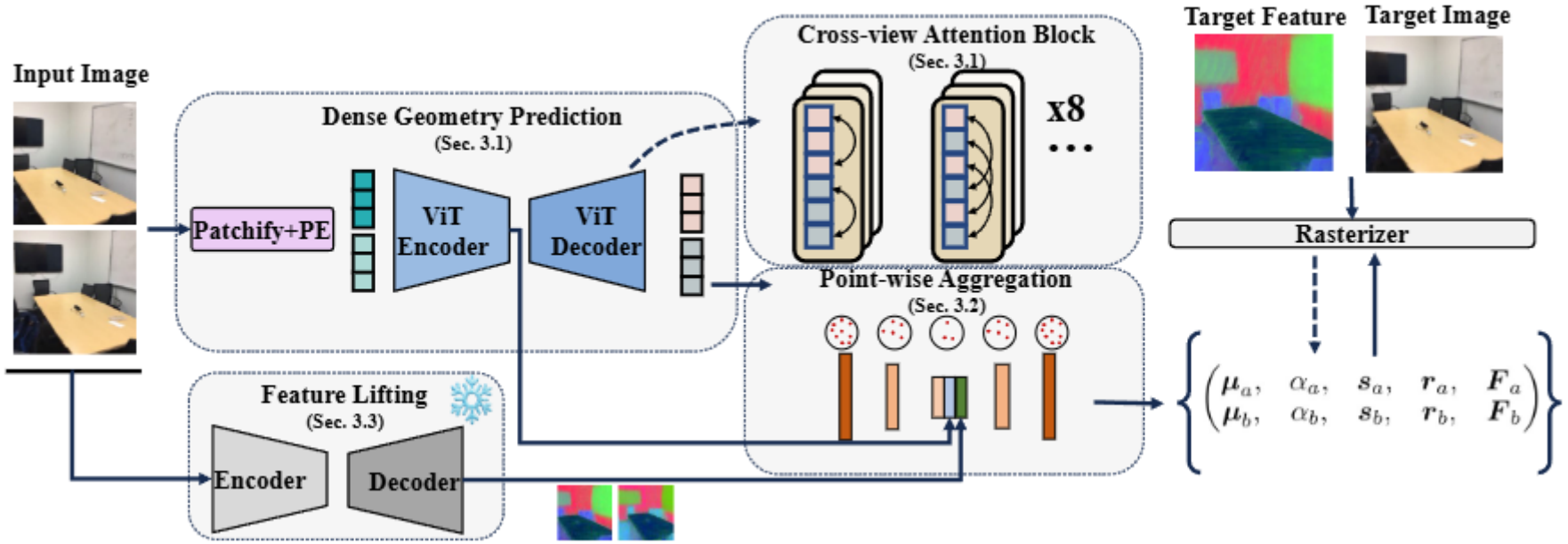
多层级余弦相似度损失 $\mathcal{L}_{sem} = \sum_{scale\in\{1,2,3\}}\left(1 - \frac{F_{pred}^{scale}\cdot F_{GT}^{scale}}{\left|F_{pred}^{scale}\right|\left|F_{GT}^{scale}\right|}\right)$

由渲染出的系数图 W 与全局码本 S 相乘得到: $F_{pred} = W \cdot S$

Feed-Forward-based Methods

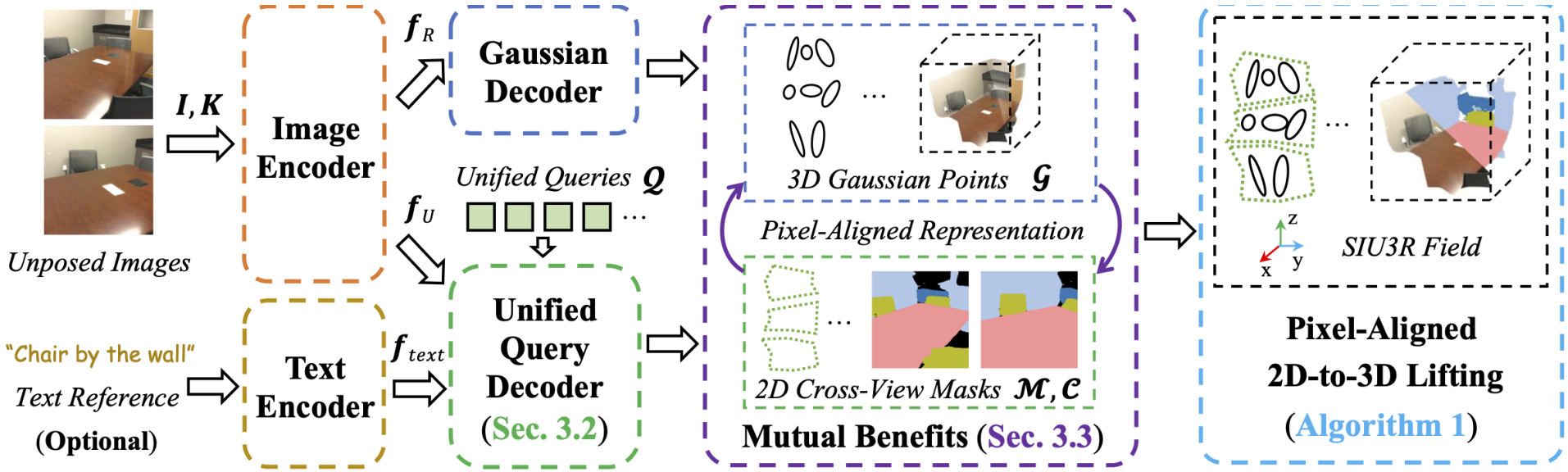
Large Spatial Model:End-to-end Unposed Images to Semantic 3D

2024 NeurIPS



SIU3R: Simultaneous Scene Understanding and 3D Reconstruction Beyond Feature Alignment

2025 NeurIPS



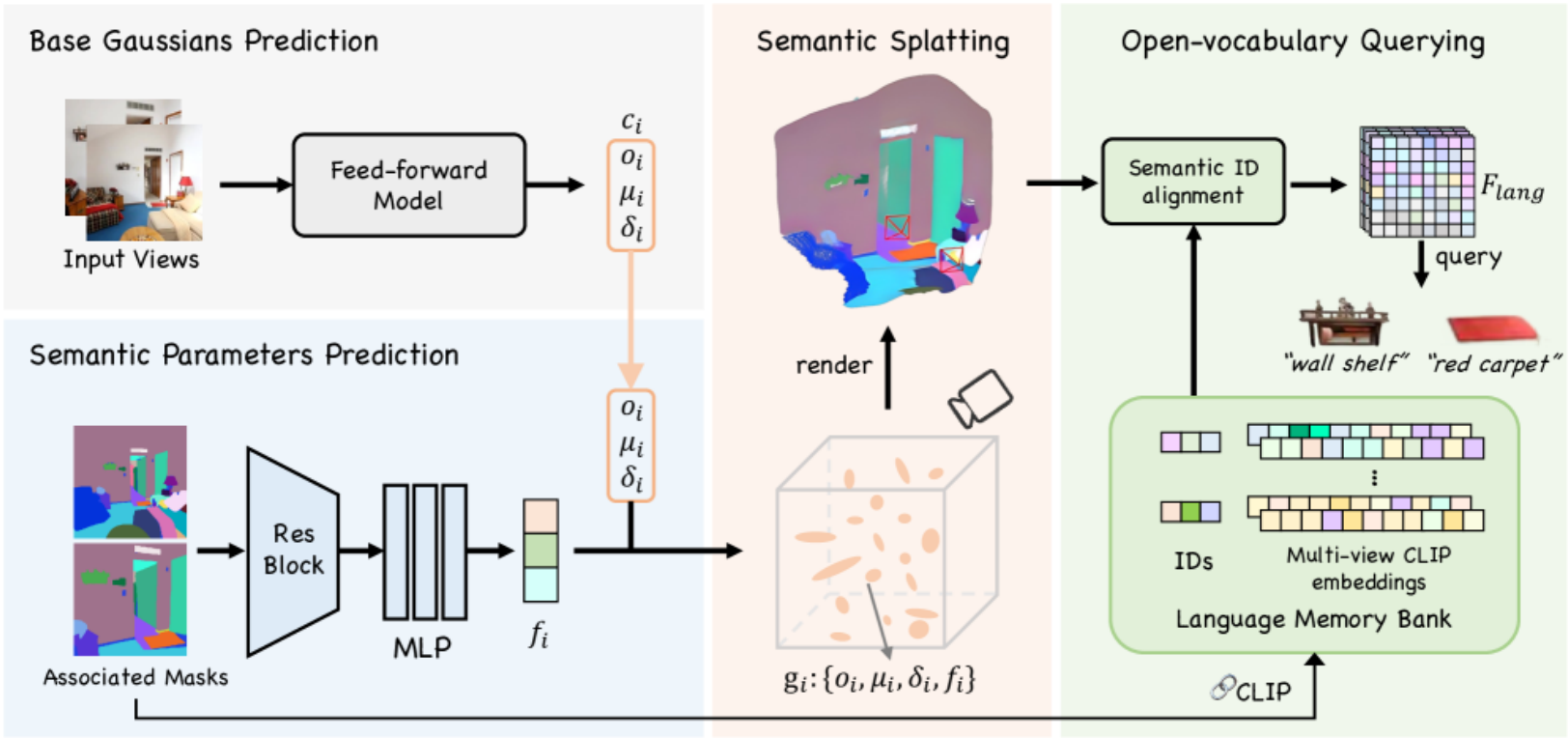
SLGaussian: Fast Language Gaussian Splatting in Sparse Views

2025 ACM MM

MVSplat

索引 $f_i \rightarrow ID$

$$\mathcal{L} = \text{MSE}(\hat{F}, M) = \frac{1}{N}\sum_{v=1}^{N} \|\hat{F}(v) - M(v)\|^2$$



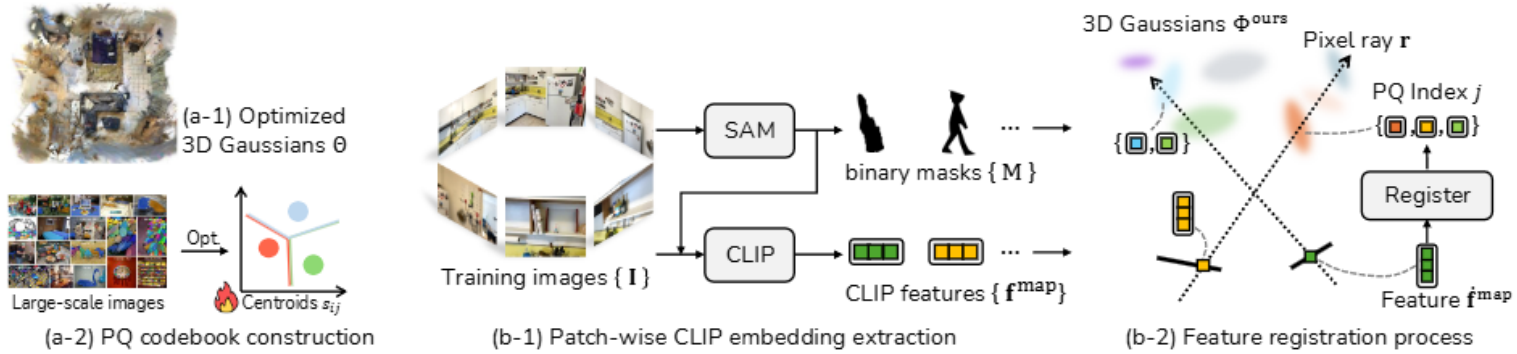
$\hat{F}(v)$： 对高斯点的语义参数f_i进行 Splatting 渲染得到

$M(v)$:视频跟踪算法预先生成的、在多视角下具有空间一致性的语义标签 ID

Dr. Splat: Directly Referring 3D Gaussian Splattingvia Direct Language Embedding Registration

2025 CVPR

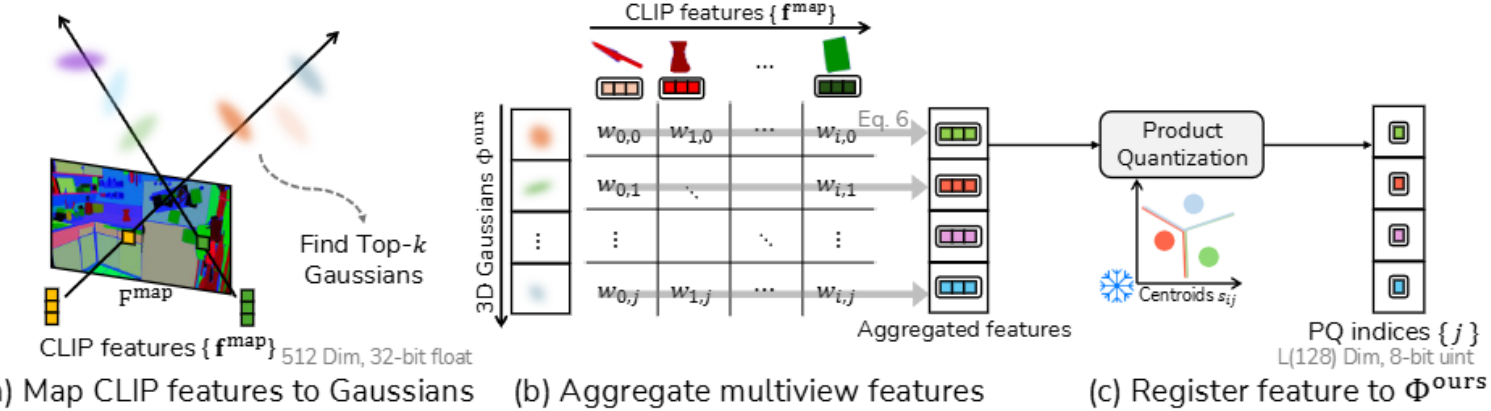
码本 $512 \rightarrow 4_{128}(K_{256})$ Top_K



$w_i(I,r) = T_i(I,r) \cdot \tilde{\alpha}_i(I,r)$

T_i:透射率表示光线到达该点前未被遮挡的比例

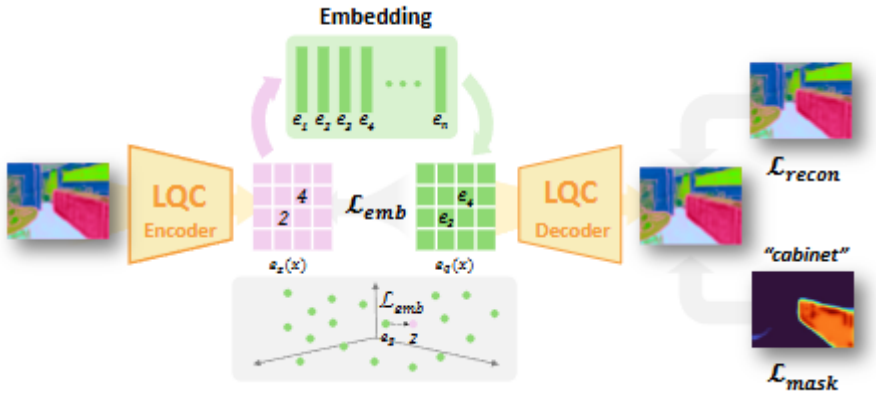
$\tilde{\alpha}_i$:有效不透明度，结合了高斯点的透明度及其在像素空间中的投影影响

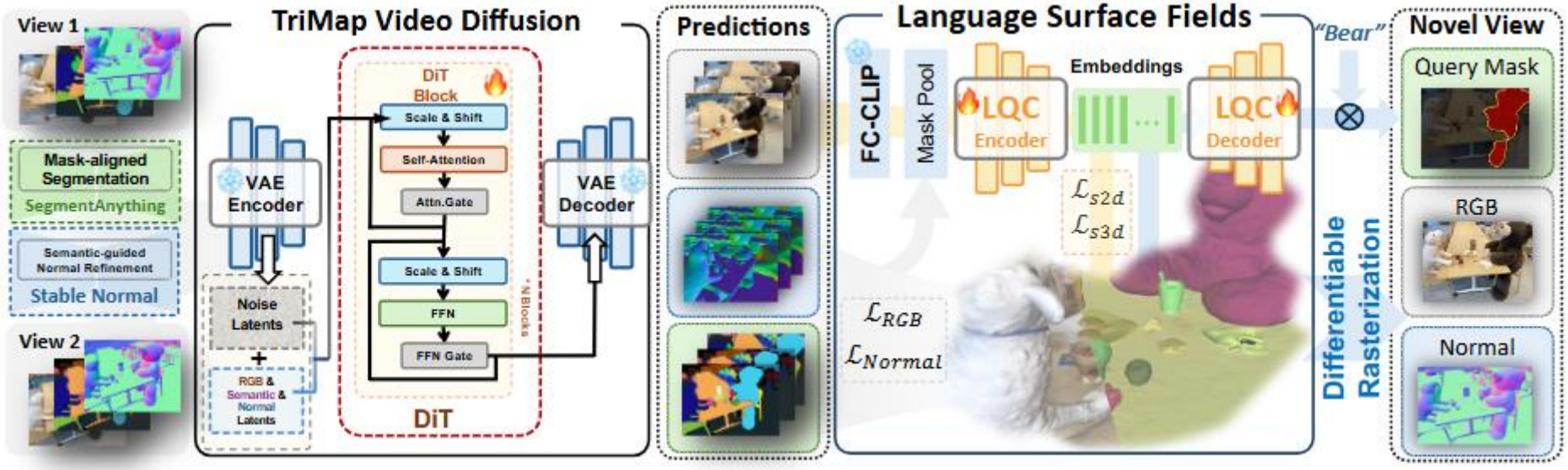


LangScene-X: Reconstruct Generalizable 3D Language-Embedded Scenes with TriMap Video Diffusion

2025 ICCV

码本 $f_i \rightarrow ID$





EmbodiedSplat: Online Feed-Forward Semantic 3DGSfor Open-Vocabulary 3D Scene Understanding

2026 CVPR

码本 $f_i \rightarrow ID$