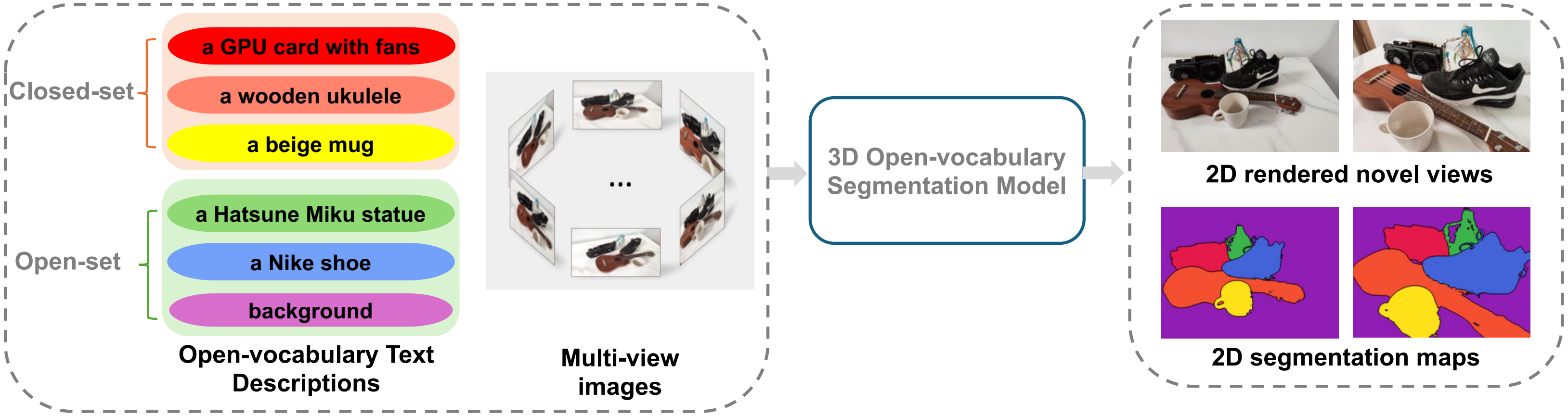
# econSG: Efficient and Multi-view Consistent Open-Vocabulary 3D Semantic Gaussians

Can Zhang    Gim Hee Lee

2 Apr 2025

# 3D Open-vocabulary Segmentation

- Produce accurate object boundaries for open-world classes in the 3D scene without requiring any segmentation annotations during training.

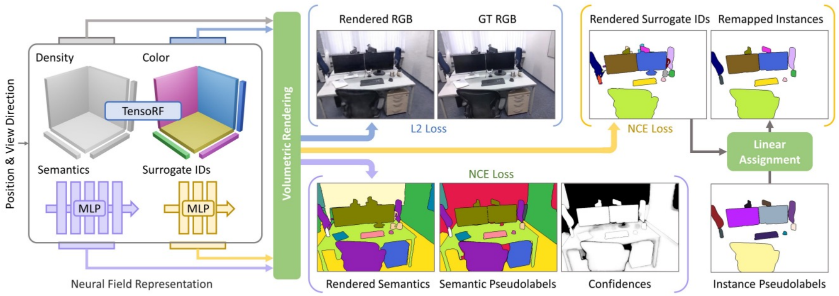



Input:
- Multi-view images
- Open-vocabulary text descriptions

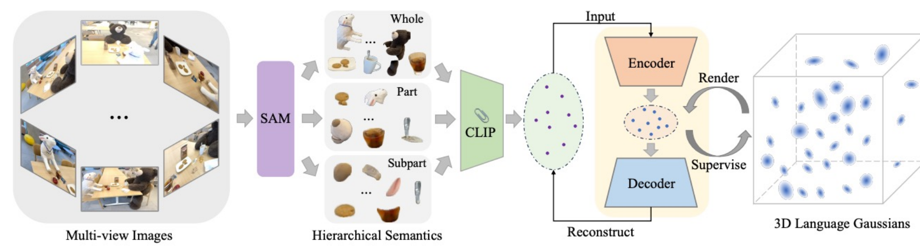
Output:
- Rendered images from novel views.
- Segmentation maps

**Note:** Train on closed-set classes and inference on both open-set and closed-set classes

# Previous 3D Segmentation Approaches

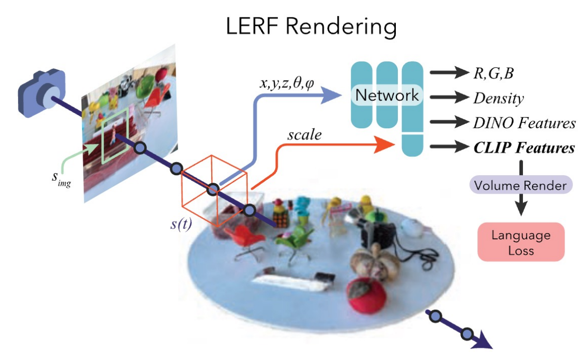



Panoptic Lifting [CVPR 2023]




LangSplat [CVPR 2024]

- Main issues:
  - Suffering from poor quality on in-the-wild scenes (e.g. Panoptic Lifting)
  - Dilemma of balancing open-vocabulary ability and localizing accuracy (e.g. LERF)
  - Time-consuming and inconsistent across views (e.g. LangSplat)




LERF [ICCV 2023]

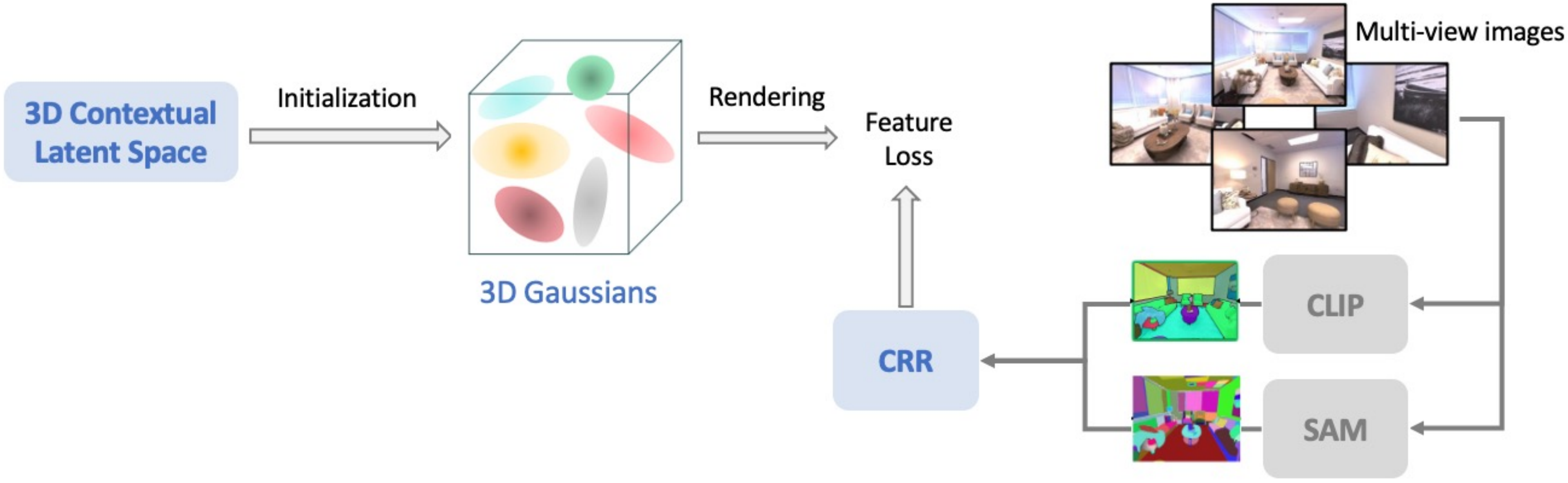
# Our Approach: econSG

- We propose a 2D-3D mapping strategy with the zero-shot open-vocabulary prompted paradigm for building a 3D semantic field in 3D Gaussian Splatting.

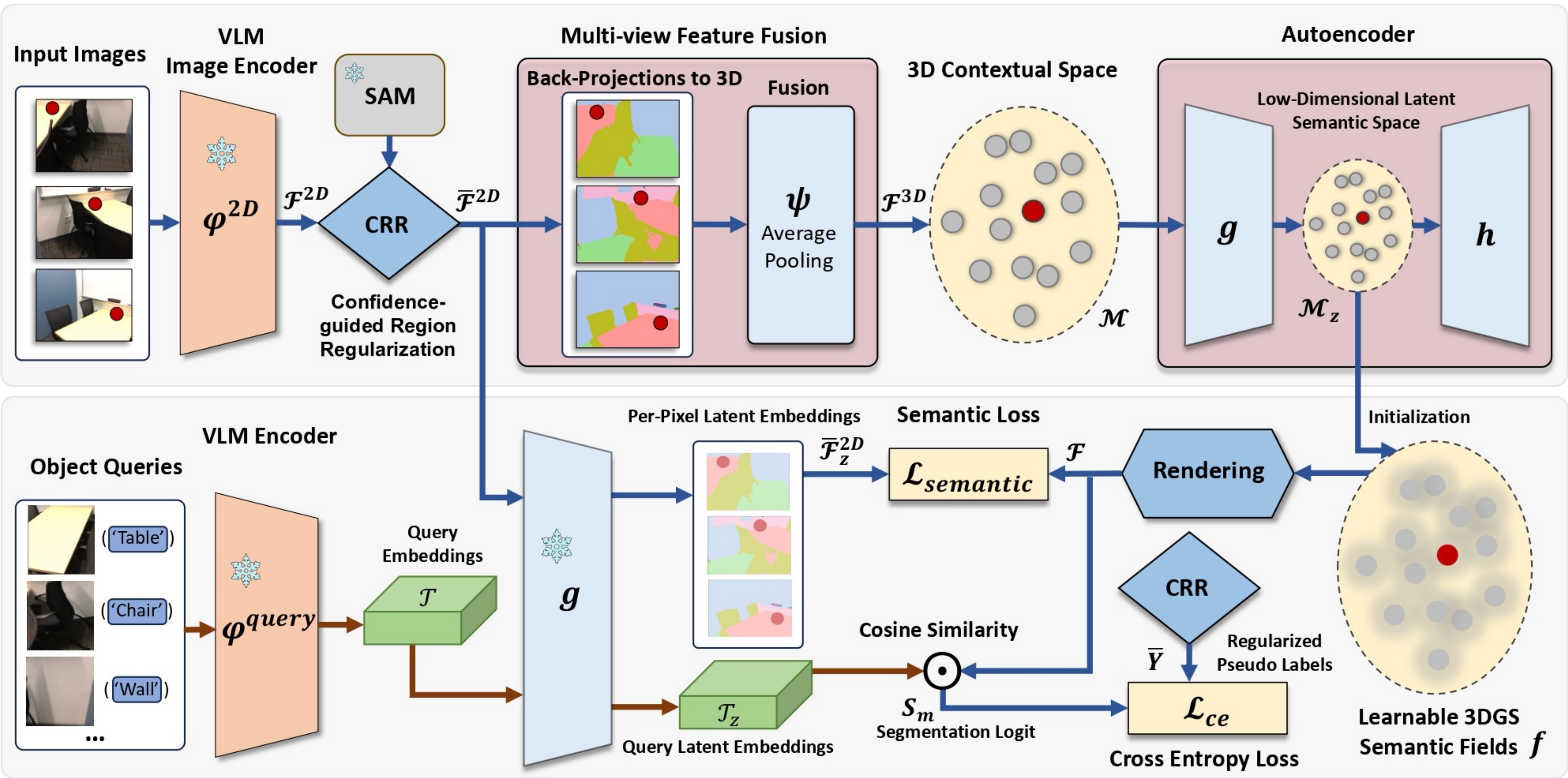
- Our approach is able to simultaneously achieve:
  - accuracy
  - efficiency
  - the grounding of open-vocabulary level semantics
  - natural modality interactions

# Our Approach: econSG

- Our model consists of three key components:
  - 3D Gaussian Splatting (3DGS) as explicit 3D representation for building open-vocabulary semantic fields
  - Confidence-guided region regularization (CRR) for semantic mask refinement
  - 3D contextual latent space for initialization

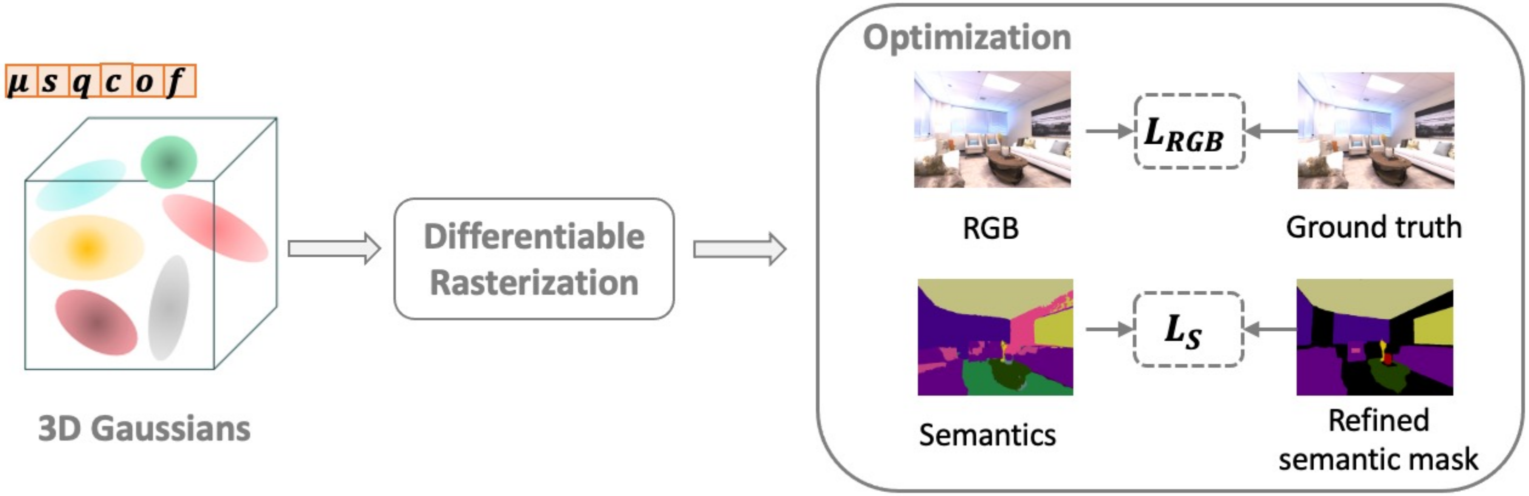


# econSG: Our Network Architecture



# 3D Gaussian Splatting (3DGS)

- Learnable parameters of each 3D Gaussian:
  - $\mu$ : 3D position
  - $s$ : Scaling factor
  - $q$ : Rotation quaternion
  - $c$ : RGB color
  - $o$ : Opacity value
  - $f$ : Semantic embedding (Additional)




Rendering equation:

$$I = R(\mu, \Sigma, c, o; p_{cam})$$

Where the result image $I$ is rendered from a specific camera pose $p_{cam}$, using the differentiable rasterization $R$.

# Confidence-guided Region Regularization (CRR)

- **Problems:**
  1. Extracting pixel-aligned CLIP features (trained for image level) from image crops shows ambiguity around object boundary
  2. SAM provides good object boundaries but show ambiguity due to a lack of contextual information and scale

- **Solution:**

  We propose to generate semantics with less ambiguities in object boundaries, contextual information and scale

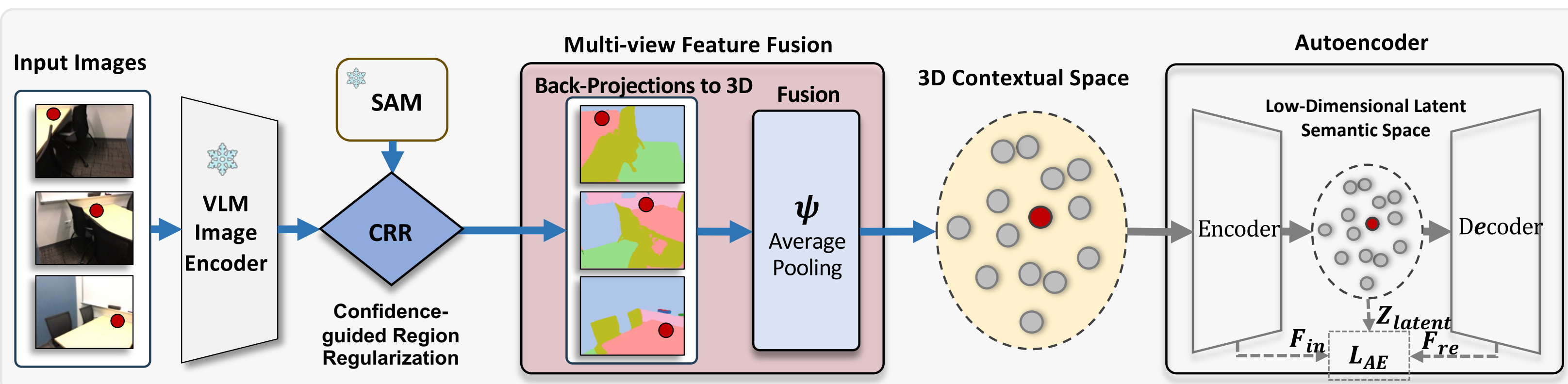
Images (a) image crops (b) SAM Crops

# Confidence-guided Region Regularization (CRR)

- Given pixel-level CLIP features and SAM masks,
  a) Select high-confidence semantic regions across all views with a higher confidence threshold $\tau_1$ .
  b) Assign semantic labels to SAM masks by the majority voting of labels in high-confidence regions within each mask.
  c) Generate semantics by:
     1) Overlapped regions: retain the semantic labels from the high-confidence maps
     2) Outside: assign with SAM labels
  d) Regularize the boundary of semantics by filtering with a lower confidence threshold $\tau_2$ .

# 3D Contextual Latent Space

- **Multi-view Feature Fusion**:
  1. Given multi-view images, extract pixel-level semantics with VLM and generate object masks with SAM.
  2. Refine 2D semantics with CRR module.
  3. Back-project 2D semantics to 3D space and fuse multi-view features to construct 3D contextual space



# 3D Contextual Latent Space

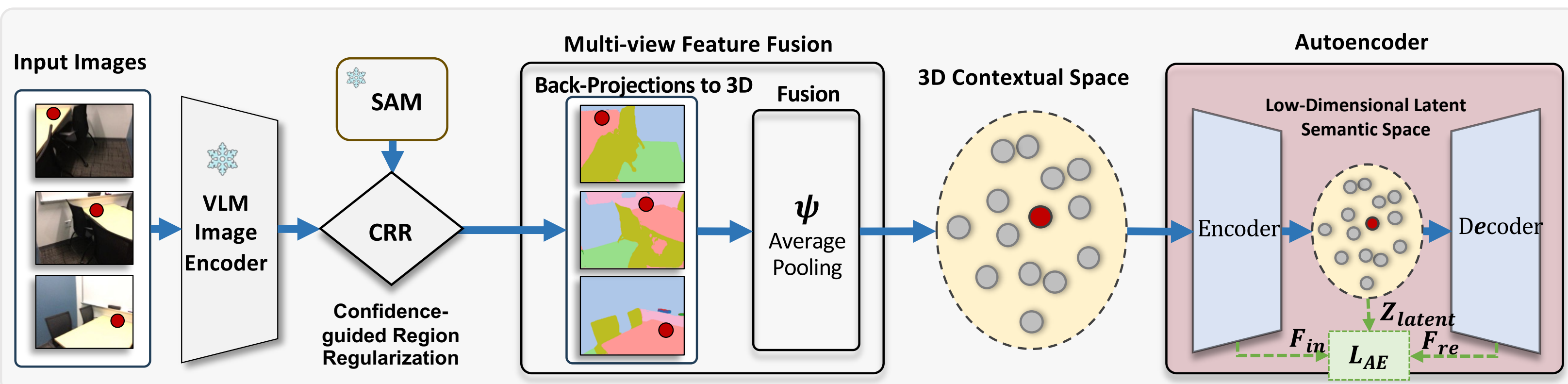
- **Autoencoder**:
  1. Train an autoencoder with the 3D contextual space.

     ❖ Reconstruction loss:

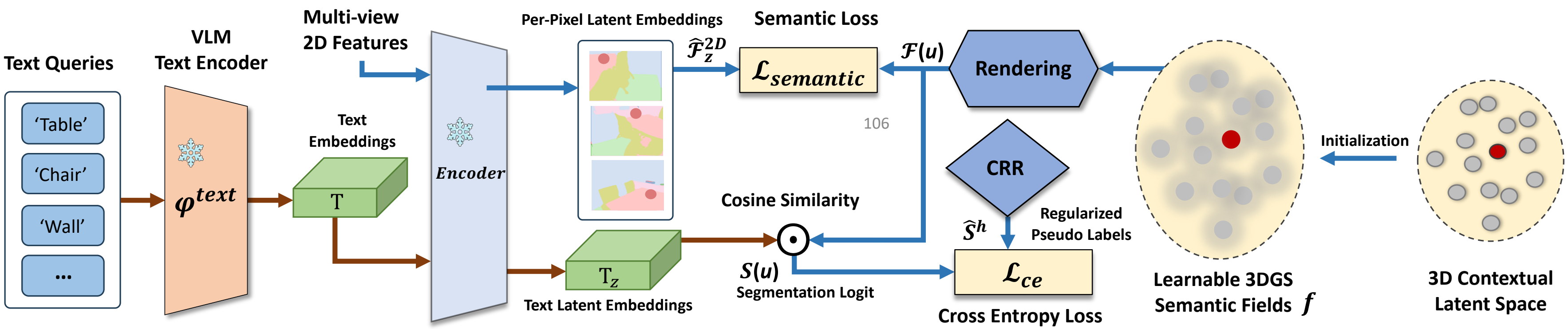
     Semantic labels

     $$L_{AE} = L_{l2}(F_{in}, F_{re}) + L_{ce}(F_{re}, \hat{y}) + L_{ce}(z_{latent}, \boxed{\hat{y}})$$

  2. Map high-dimensional 3D contextual space into low-dimensional latent semantic space with the encoder.



# 3D Gaussian Splatting for Semantic Fields




- Initialize the semantic embedding $f$ in each 3D Gaussian with the low-dimensional 3D contextual latent space.
- We optimize the 3DGS semantic fields with:

$$\mathcal{L}_{semantic} = \sum_{u=1}^{U} Dist(\mathcal{F}(u), \hat{\mathcal{F}}_z^{2D}(u)),\ \mathcal{L}_{ce} = CE(S(u), \hat{S}^h)$$

  - $Dist(\cdot,\cdot)$ denotes the distance function.
  - $S(u) = cos < \mathcal{F}(u), \mathcal{T}_z >$ denotes the segmentation logit at pixel $u$.

Final Losses: $\mathcal{L} = \boxed{\mathcal{L}_{color}} + \lambda_{sem}\mathcal{L}_{semantic} + \lambda_{2d}\mathcal{L}_{ce}(u)$

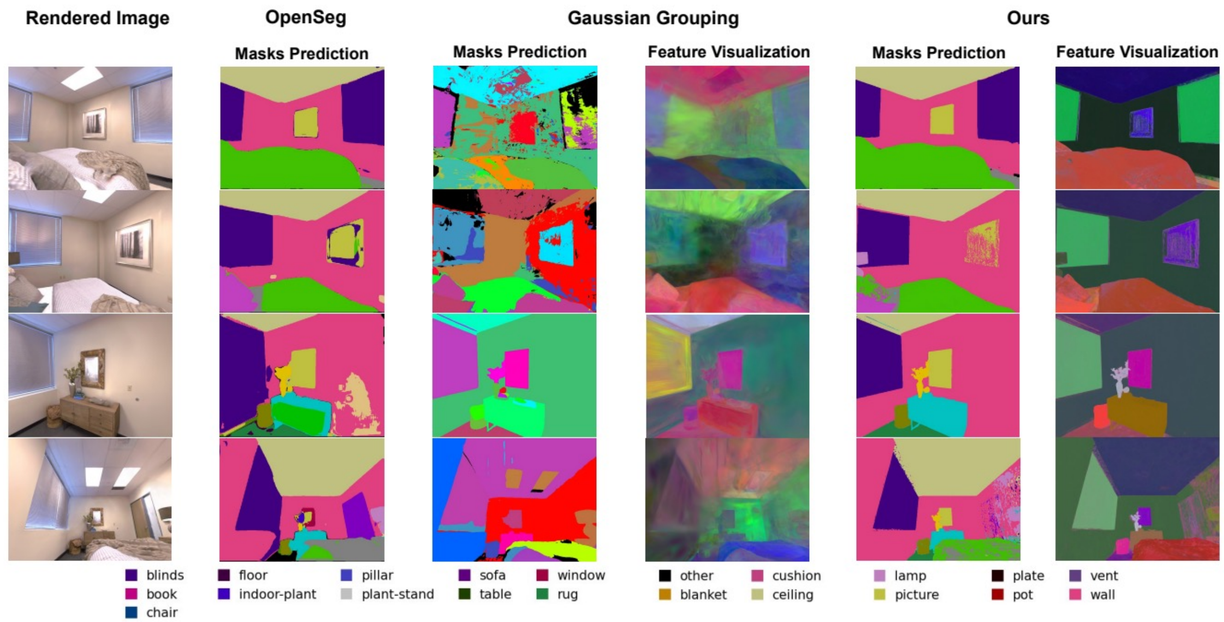
RGB fields

NUS National University of Singapore | School of Computing

# Segmentation Results of Novel Views from Scannet and Replica

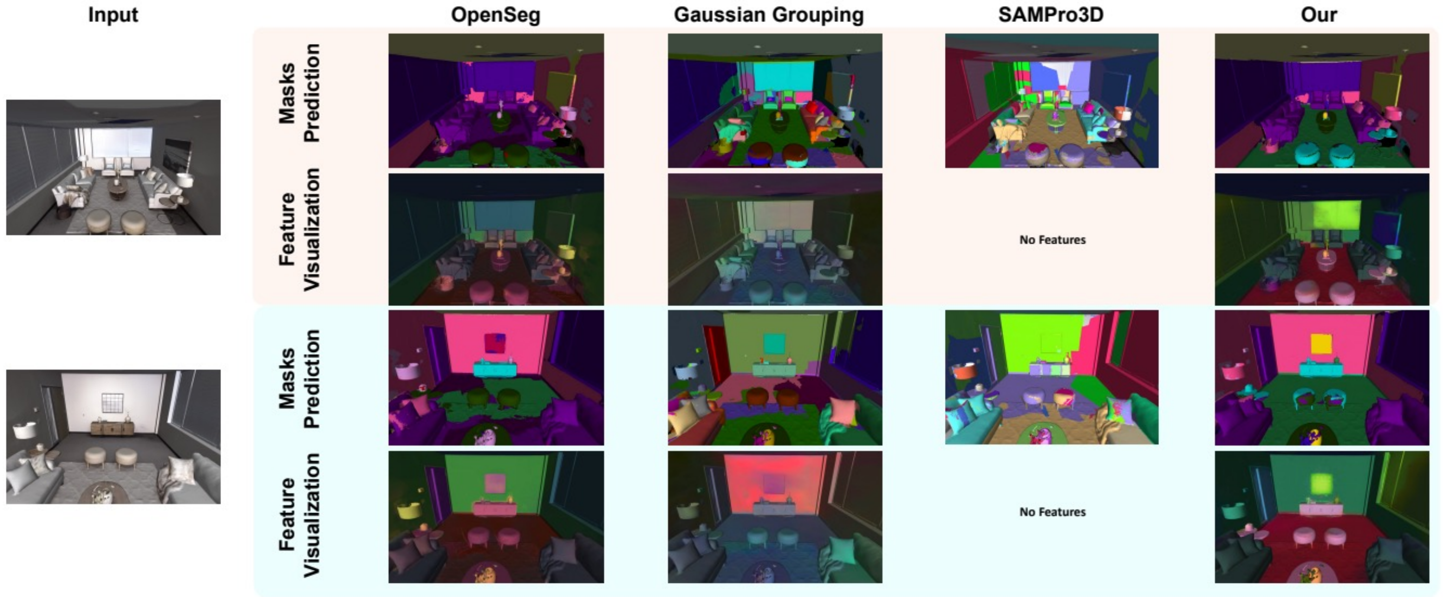
- Open-vocabulary Segmentation Comparison

| Dataset | FPS | Replica | | | | Scannet | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | sparse-view | | multi-view | | sparse-view | | multi-view | |
| | | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc |
| LERF | 0.2 | 4.312 | 17.080 | 8.285 | 22.125 | 14.059 | 38.734 | 15.349 | 40.294 |
| 3DOVS | 0.3 | 4.553 | 19.356 | 9.081 | 23.938 | 14.227 | 40.584 | 17.802 | 42.532 |
| Feature3DGS | 2.5 | 9.584 | 38.245 | 10.634 | 36.520 | 17.552 | 48.686 | 18.069 | 54.101 |
| econSG (Ours) | 156 | **25.513** | **70.716** | **33.869** | **78.564** | **39.018** | **74.805** | **48.205** | **86.178** |



# Analysis

- Analysis on the 3D Contextual Latent Space (Replica)

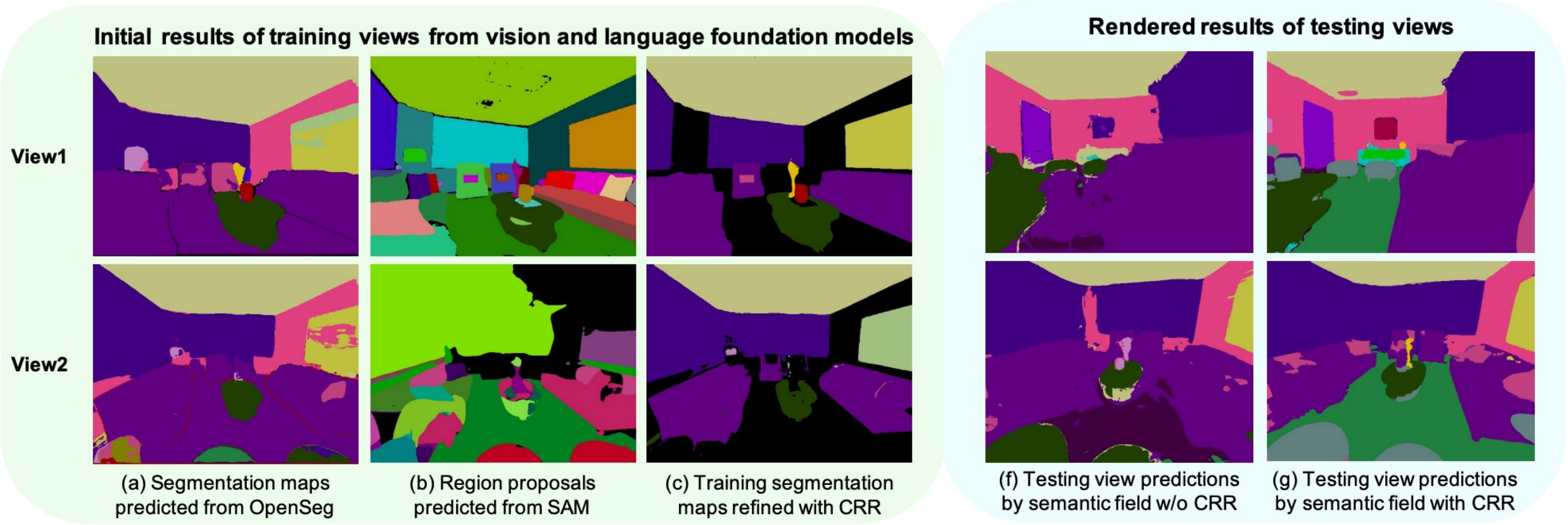



- Training efficiency analysis on the 3DOVS dataset

| Methods | LERF | 3DOVS | Langsplat | Feature3DGS | Ours | | | Ours (remove autoencoder) |
|---|---|---|---|---|---|---|---|---|
| Feature dimension | 512 | 512 | 3 | 128 | 6 | 16 | 32 | 512 |
| mIoU (%) | 27.0 | 74.0 | 82.3 | 6.7 | 91.6 | 91.8 | 91.8 | OOM |
| Training time (min) | 19.4 | 78 | 66 | 87 | 29 | 32 | 43 | OOM |
| Inference (s) | 121.4 | 6.6 | 401.9 | 6.0 | 4.9 | 5.2 | 5.3 | OOM |

# Ablation

- Ablation on Confidence-guided Region Regularization

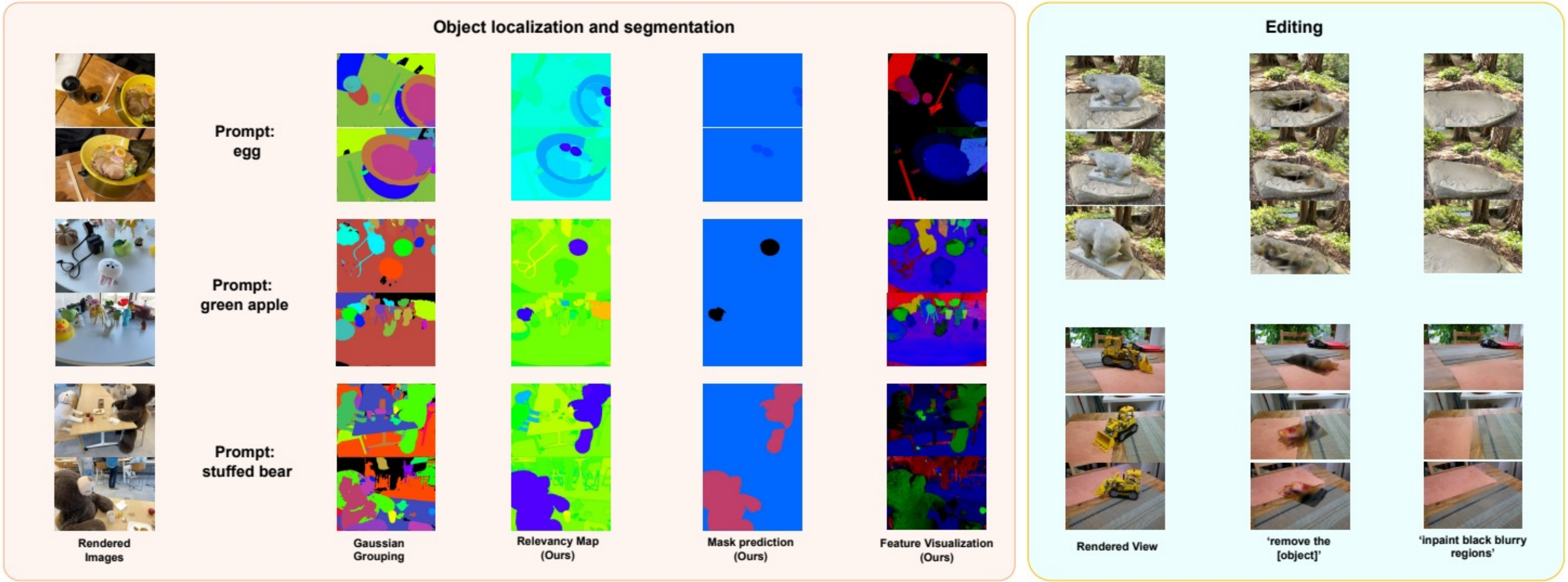



NUS National University of Singapore | School of Computing

CVRP

# Applications

- Language-guided segmentation and editing



# Thanks for Listening!