



3D Vision-Language Gaussian Splatting



Qucheng Peng



Benjamin Planche



Zhongpai Gao



Meng Zheng



Anwesa Choudhuri



Terrence Chen

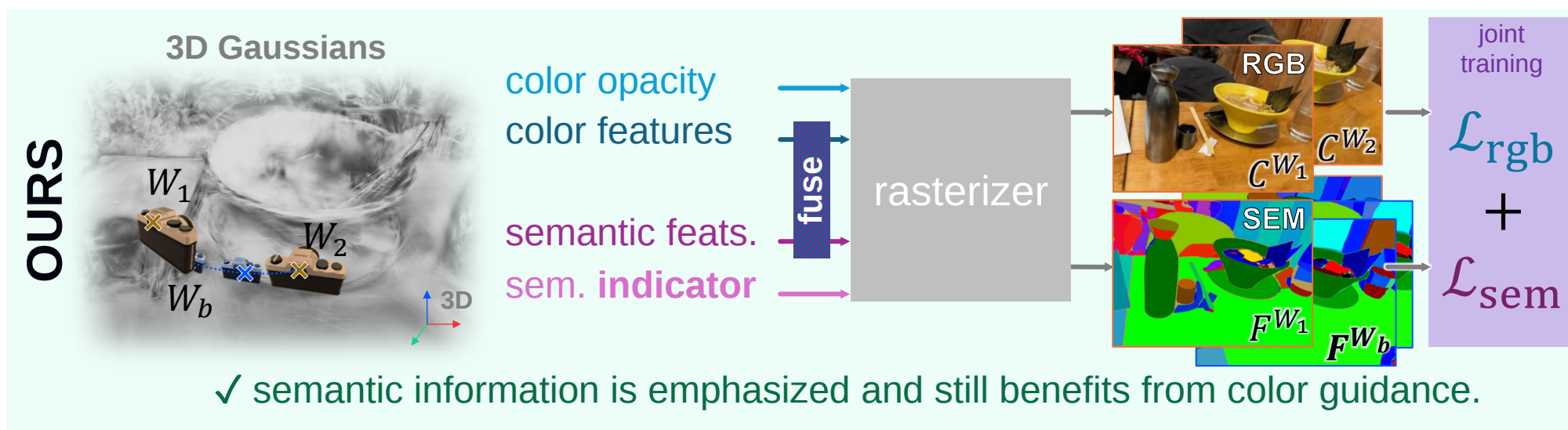
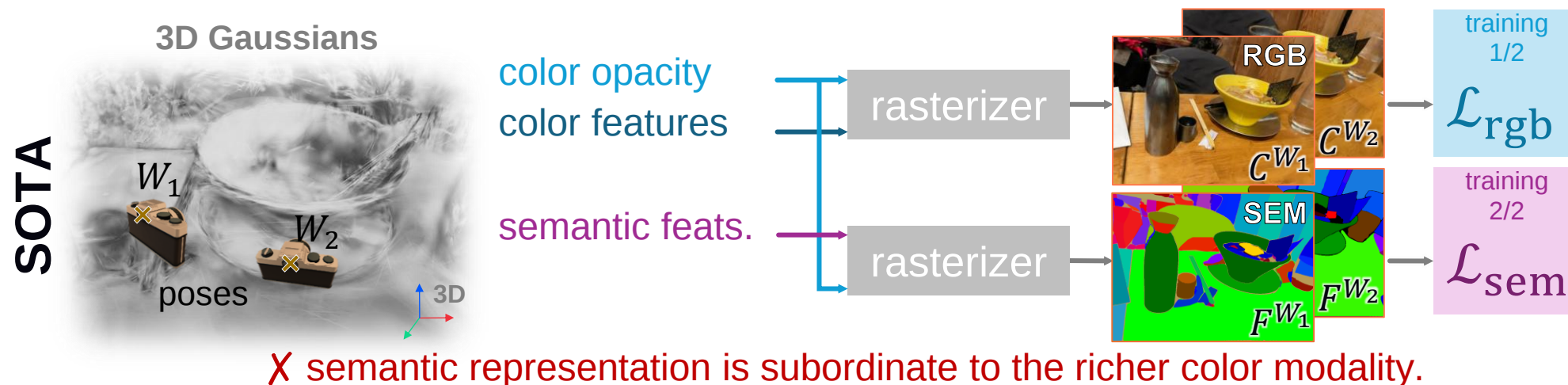


Chen Chen



Ziyang Wu





Goal: Balance the richer visual modality and the task-relevant semantic information. ✓

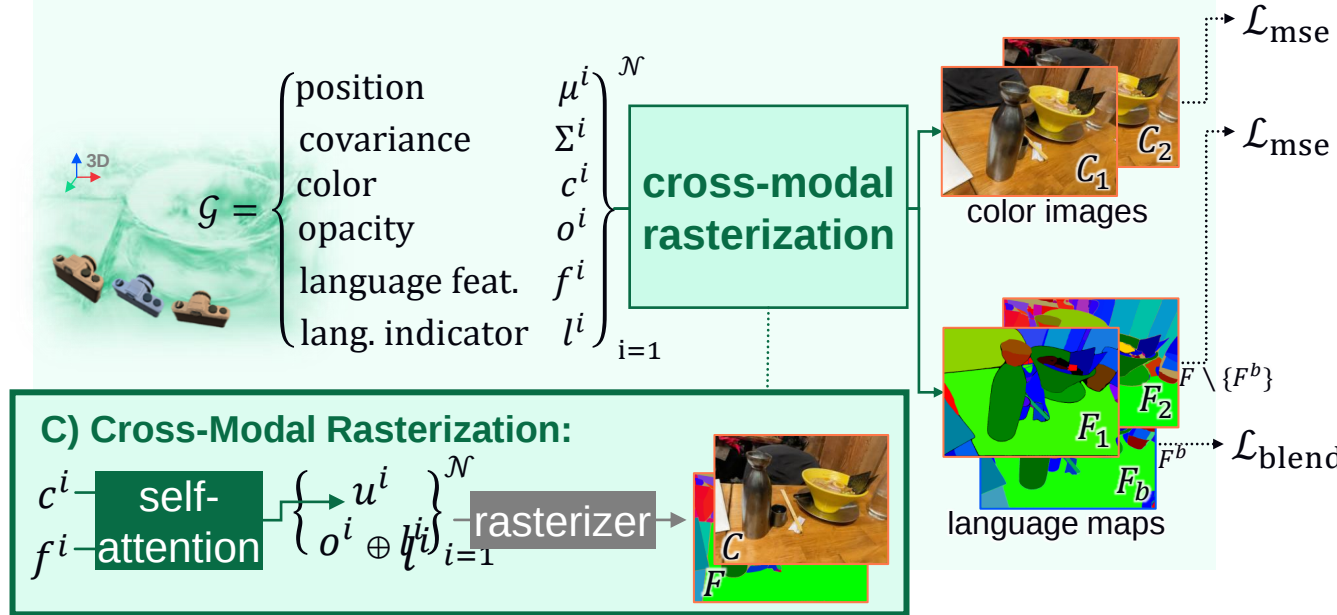
1. Semantic-aware Rasterization:

- Cross-modality fusion to exchange information between modalities before rasterization.
- Novel smoothed language indicator for the α -blending of semantic features, independently of visual translucence/reflection.

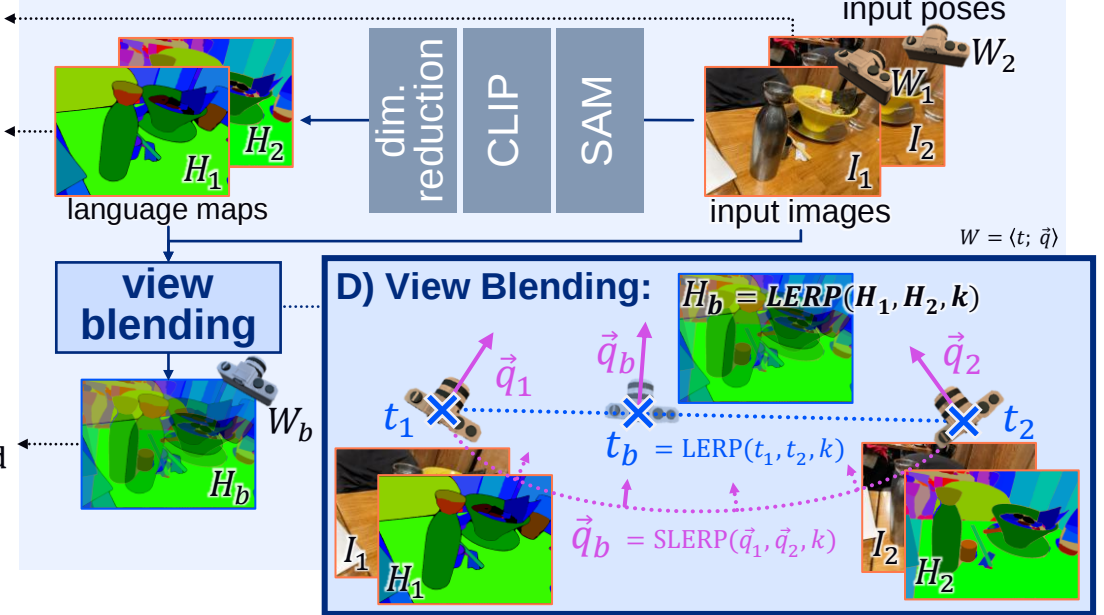
2. Semantic-aware Camera View-Blending:

- Blending-based regularization of semantic representations during training.
- Use of semantic-consistency prior to alleviate over-fitting on the color modality.

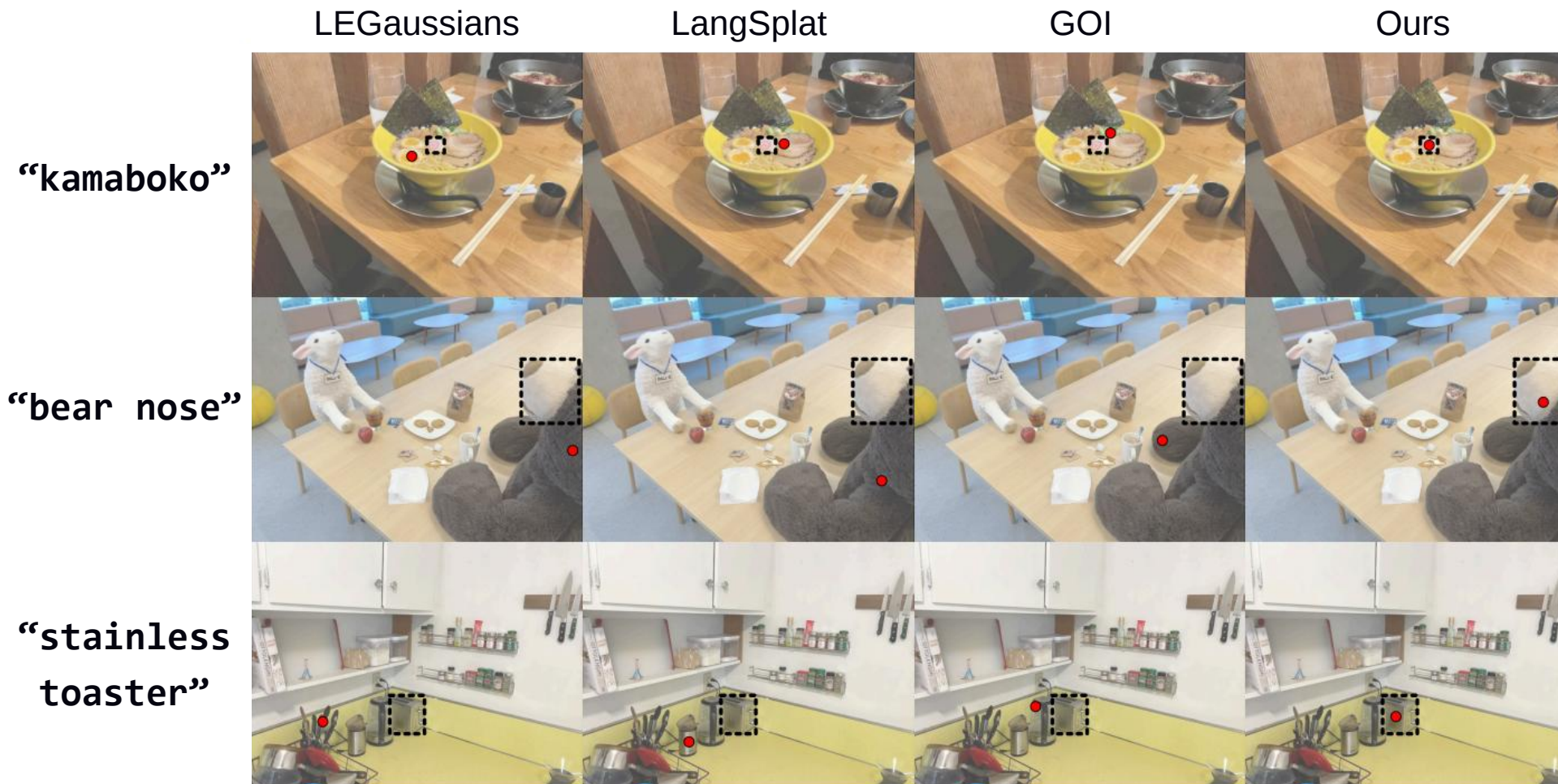
A) Multimodal Gaussian Representation



B) Data Enrichment



dataset: LERF

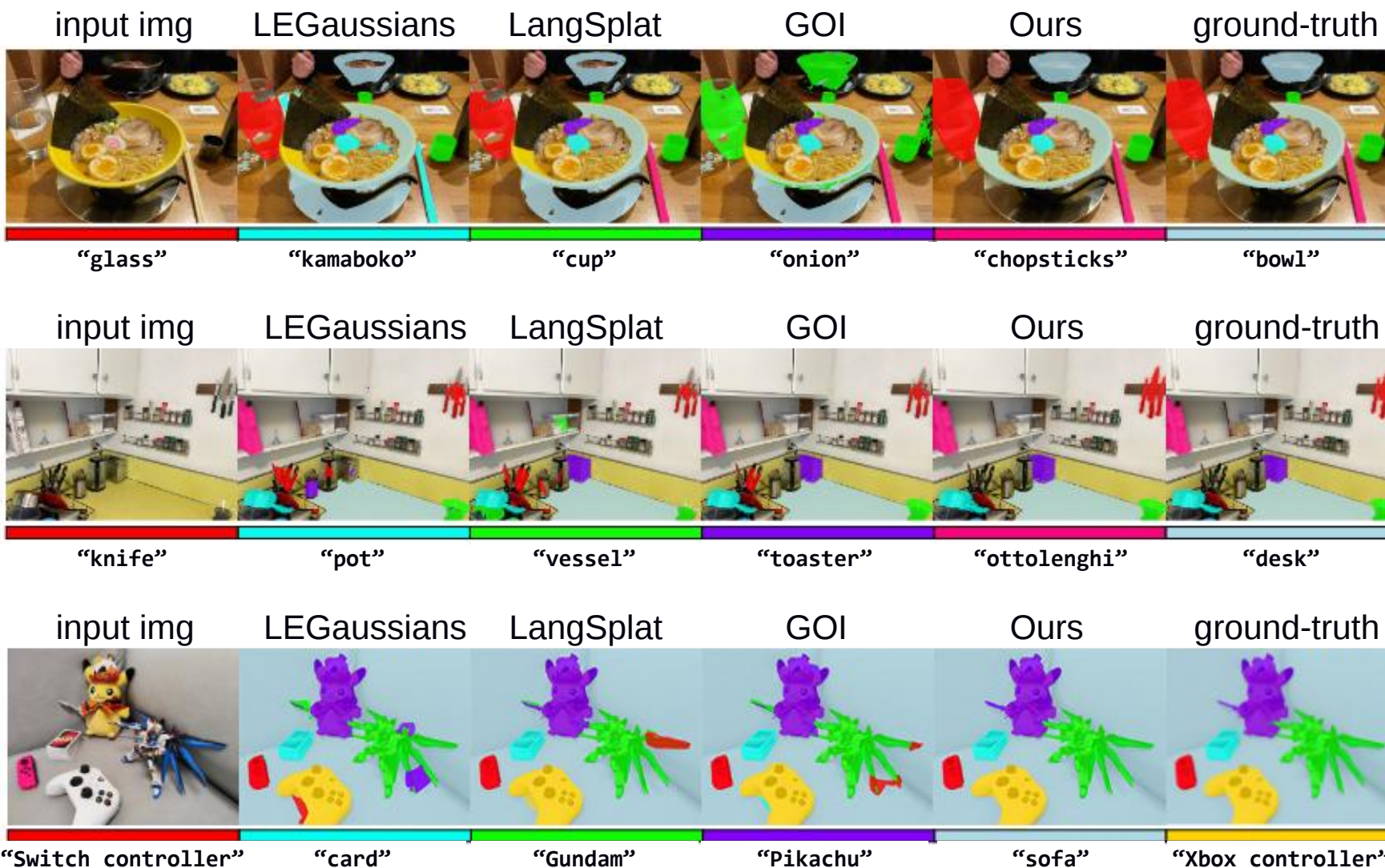


ground-truth object bounding-box

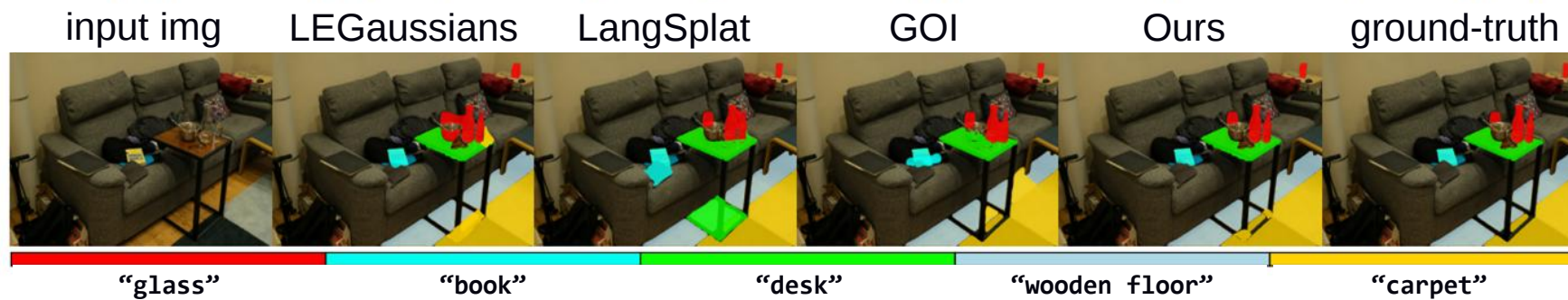


prediction (correct if within the box)

datasets: LERF & 3D-OVS



dataset: Mip-NeRF 360



dataset: LERF

prompt:

“glass”

“cup”

“knife”

“toaster”

“handle”

GT:



baseline:



ours:



Impact of the proposed additional semantic indicator on 2D relevancy maps.

dataset: LERF

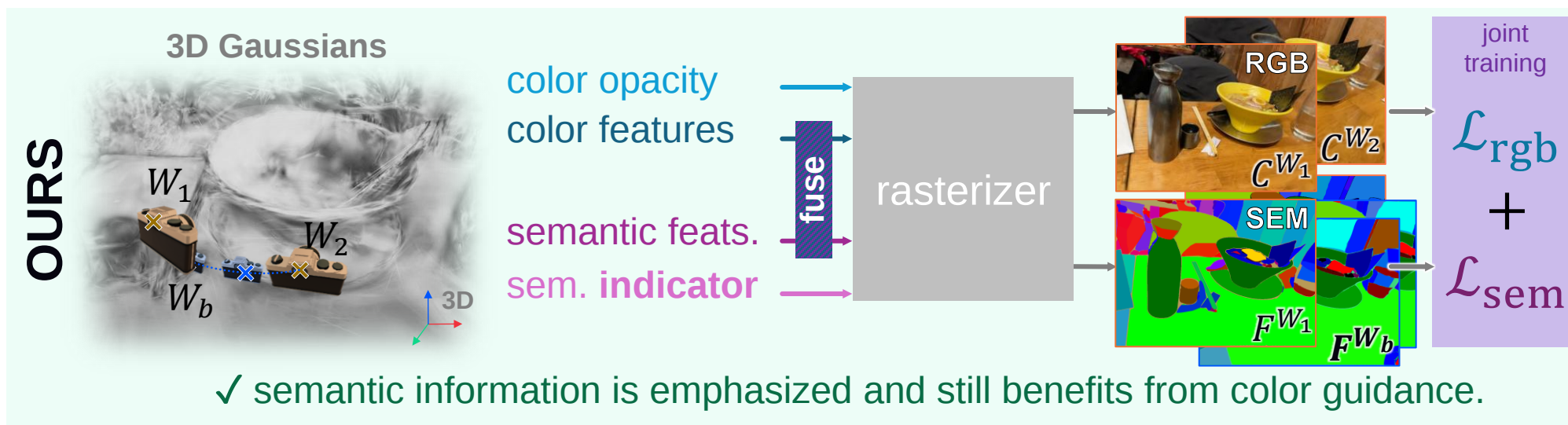
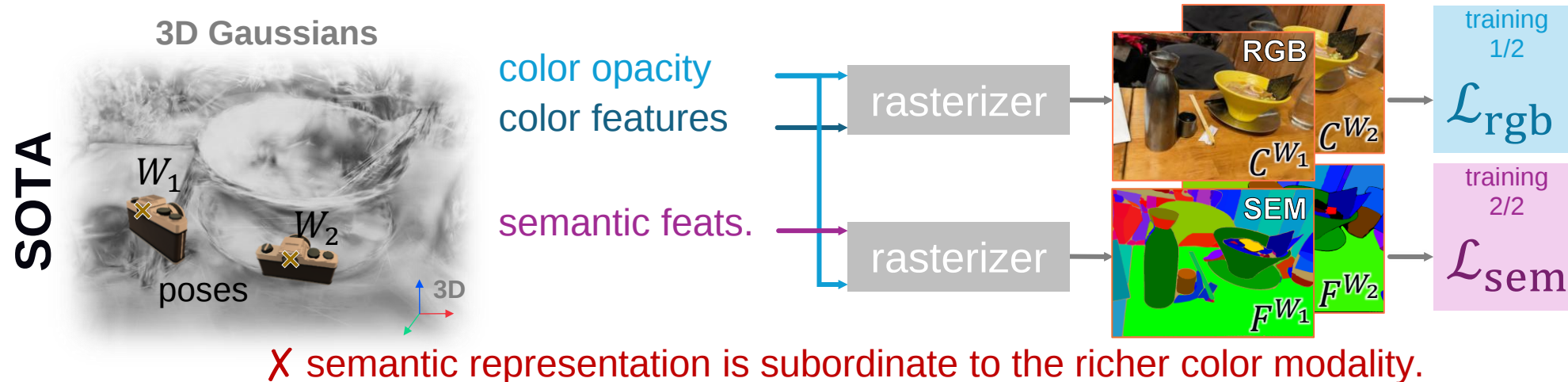
Segmentation accuracy (mIoU).

Method	Venue	ramen	teatime	figurines	kitchen	avg.
Feature-3DGS	CVPR'24	43.7	58.8	40.5	39.6	45.7
LEGaussians	CVPR'24	46.0	60.3	40.8	39.4	46.9
LangSplat	CVPR'24	51.2	<u>65.1</u>	<u>44.7</u>	<u>44.5</u>	<u>51.4</u>
GS-Grouping	ECCV'24	45.5	60.9	40.0	38.7	46.3
GOI	ACMMM'24	<u>52.6</u>	63.7	44.5	41.4	50.6
Ours		61.4	73.5	58.1	54.8	62.0

Localization accuracy (% correct).

Method	Venue	ramen	teatime	figurines	kitchen	avg.
Feature-3DGS	CVPR'24	69.8	77.2	73.4	87.6	77.0
LEGaussians	CVPR'24	67.5	75.6	75.2	90.3	77.2
LangSplat	CVPR'24	73.2	88.1	80.4	<u>95.5</u>	84.3
GS-Grouping	ECCV'24	68.6	75.0	74.3	88.2	76.5
GOI	ACMMM'24	75.5	88.6	82.9	90.4	84.4
FMGS	IJCV'24	<u>90.0</u>	<u>89.7</u>	<u>93.8</u>	92.6	<u>91.5</u>
Ours		92.5	95.8	97.1	98.6	96.0

3D Vision-Language Gaussian Splatting / take-away



- Cross-modal rasterizer focusing on language feature, for downstream accuracy.
- View-blending-based regularization to alleviate over-fitting on the color modality. ✓

Leading
CHANGE