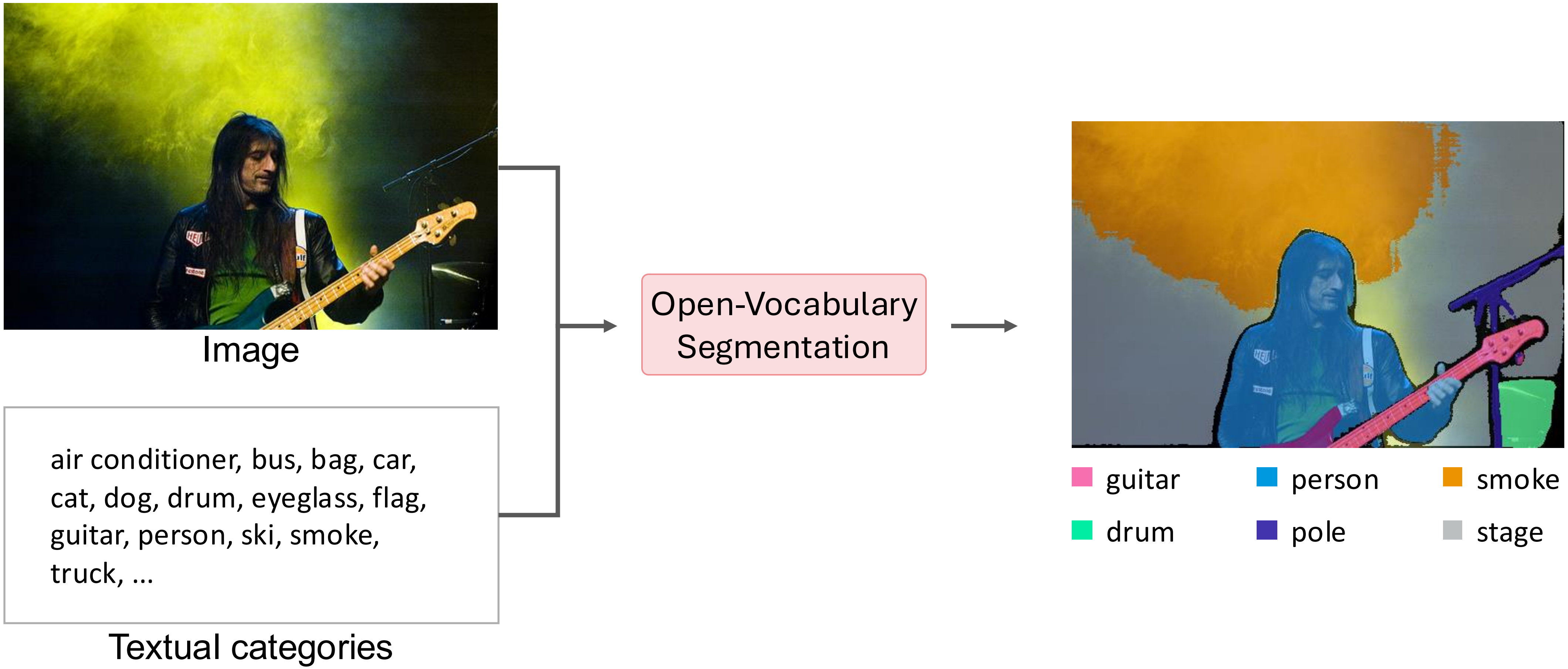
ReME: A Data-Centric Framework for Training-Free Open-Vocabulary Segmentation

Xiwei Xuan, Ziquan Deng, and Kwan-Liu Ma
University of California, Davis

ICCV HONOLULU HAWAII
OCT 19-23, 2025

Open-Vocabulary Segmentation (OVS)

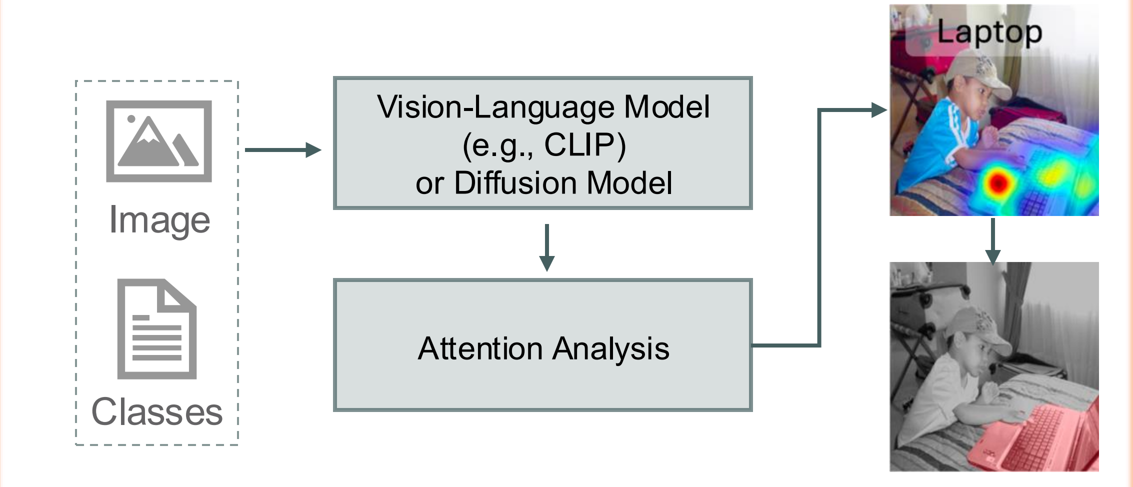
Open-vocabulary segmentation aims at segmenting images into a set of categories expressed through free-form text.



Training-Free Open-Vocabulary Segmentation (OVS)

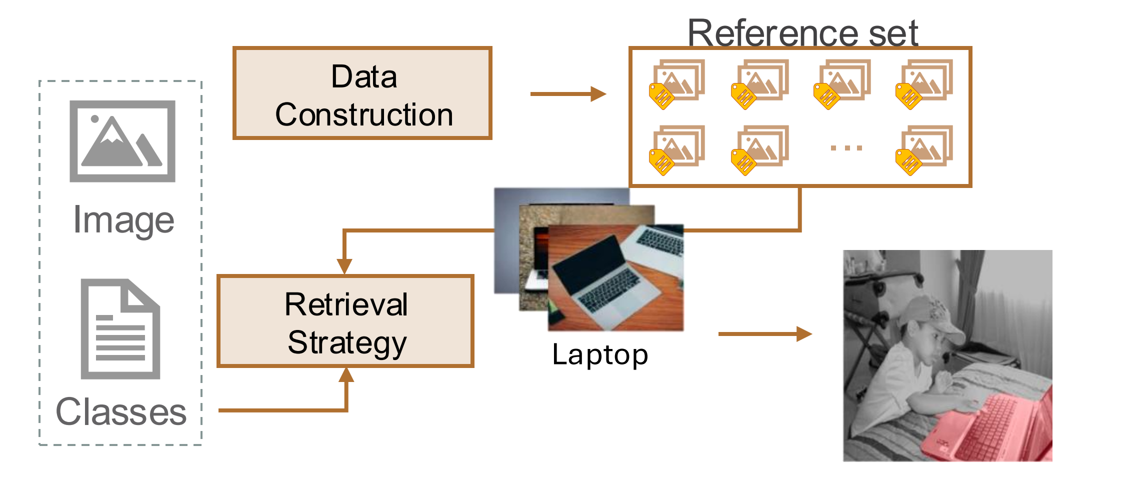
Existing Approaches

Attention-Based Approaches



- Modify the attention mechanisms of CLIP or Diffusion model
- Improve the localization capabilities with specific design

Retrieval-Based Approaches

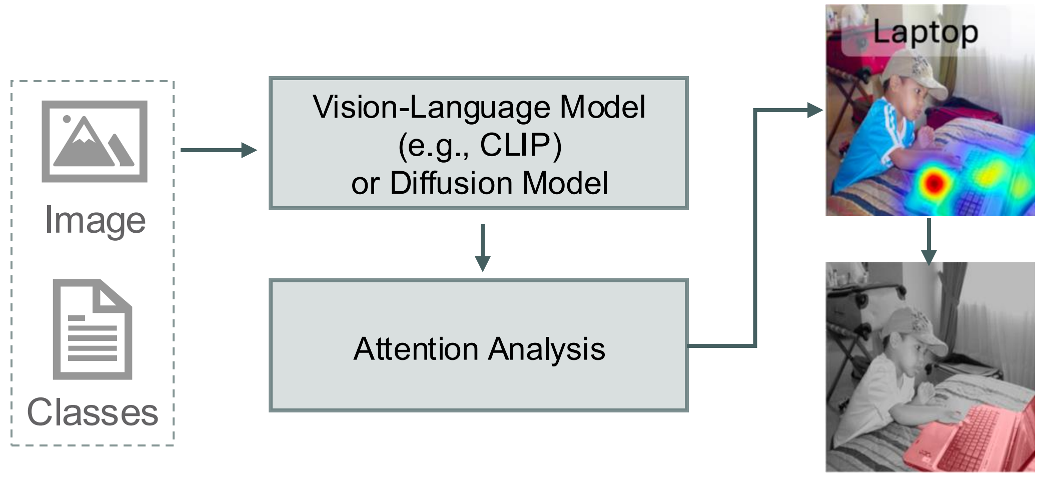


- Construct a reference set for retrieval
- Retrieve and aggregate labels for class-agnostic masks

Training-Free Open-Vocabulary Segmentation (OVS)

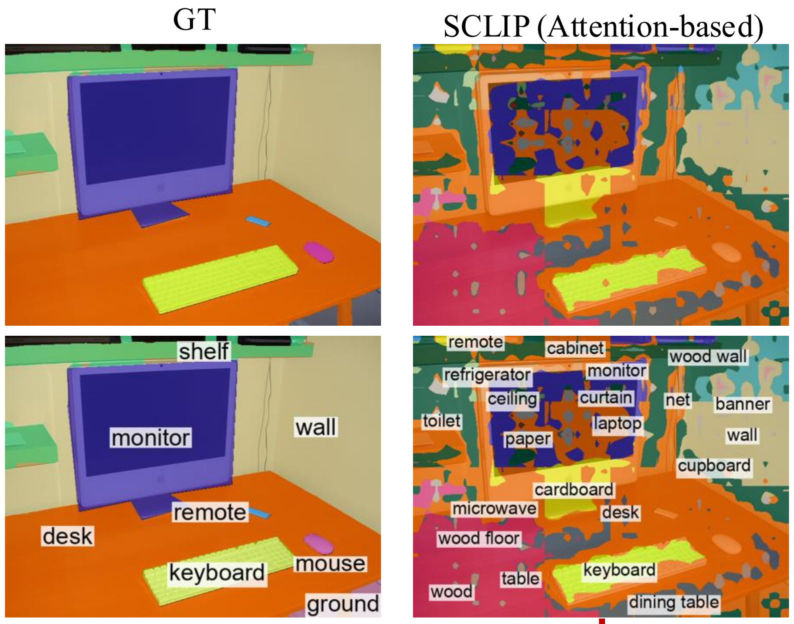
Existing Approaches

Attention-Based Approaches



- Modify the attention mechanisms of CLIP or Diffusion model
- Improve the localization capabilities with specific design

\- **Constrained by the inherent limitations of reliant models**

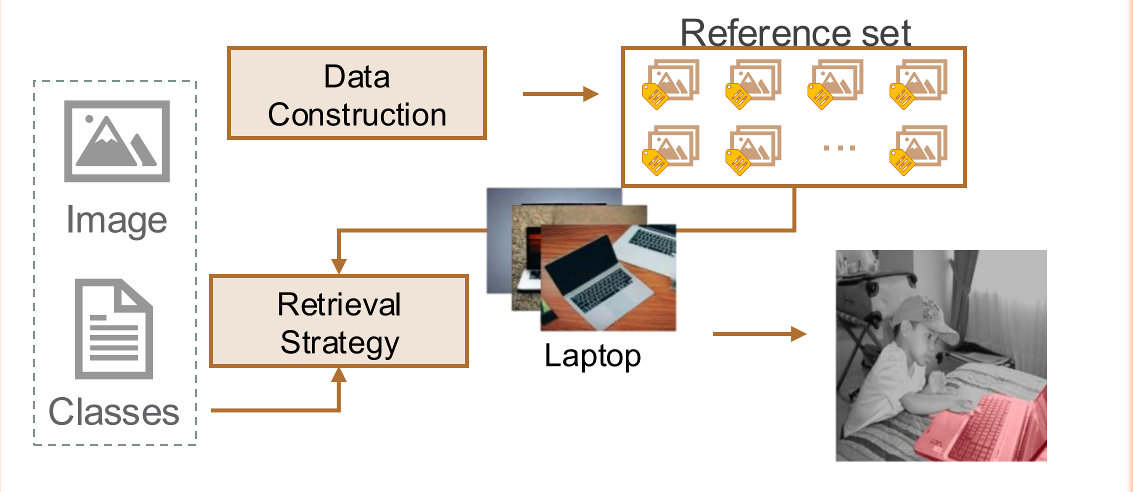


Noisy label assignments: microwave, refrigerator, curtain, ...

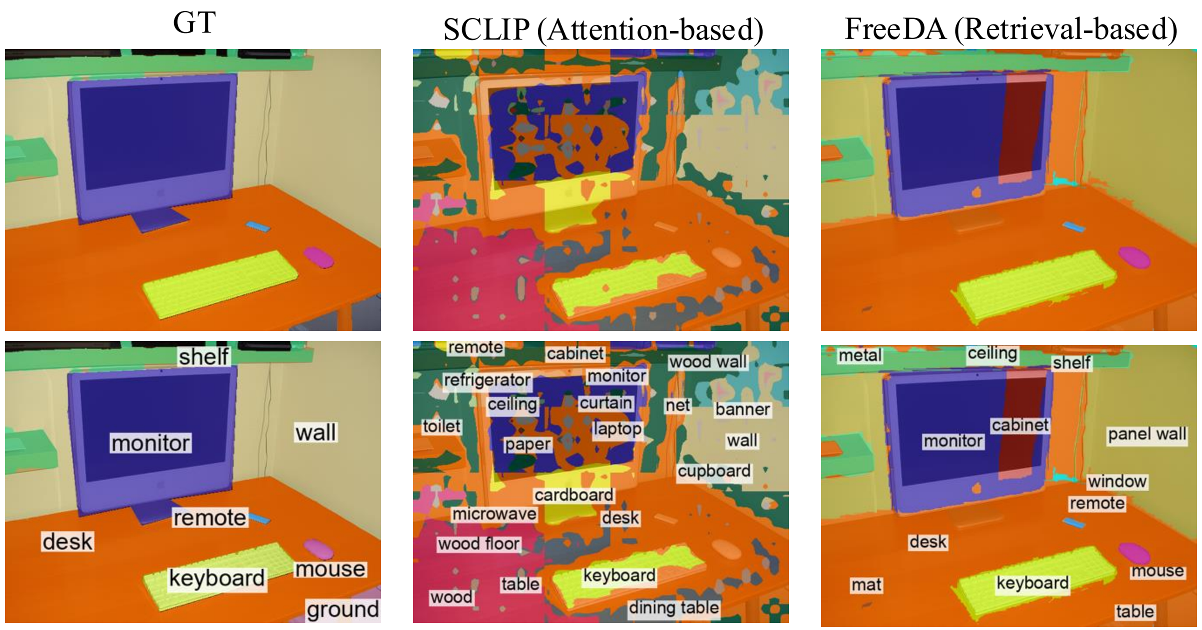
Training-Free Open-Vocabulary Segmentation (OVS)

Existing Approaches

Retrieval-Based Approaches



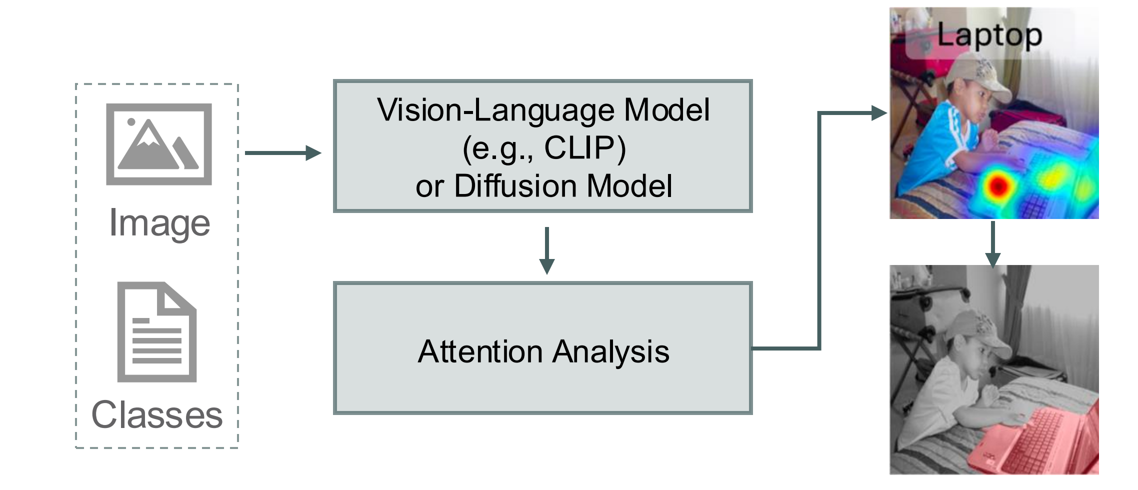
- Construct a reference set for retrieval
- Align and aggregate labels of class-agnostic masks



Training-Free Open-Vocabulary Segmentation (OVS)

Existing Approaches

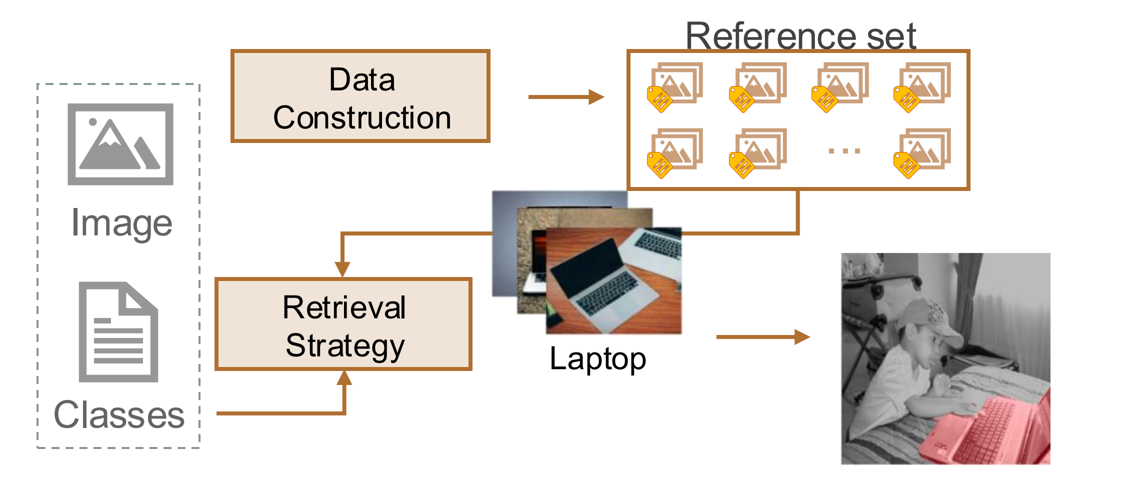
Attention-Based Approaches



- Modify the attention mechanisms of CLIP or Diffusion model
- Improve the localization capabilities with specific design

- **Constrained by the inherent limitations of reliant models**

Retrieval-Based Approaches



- Construct a reference set for retrieval
- Align and aggregate labels of class-agnostic masks

+ **Correcting VLM vulnerabilities by reference substances**
- **Overlook the fundamental data quality issues**

Preliminary Study on Data Quality

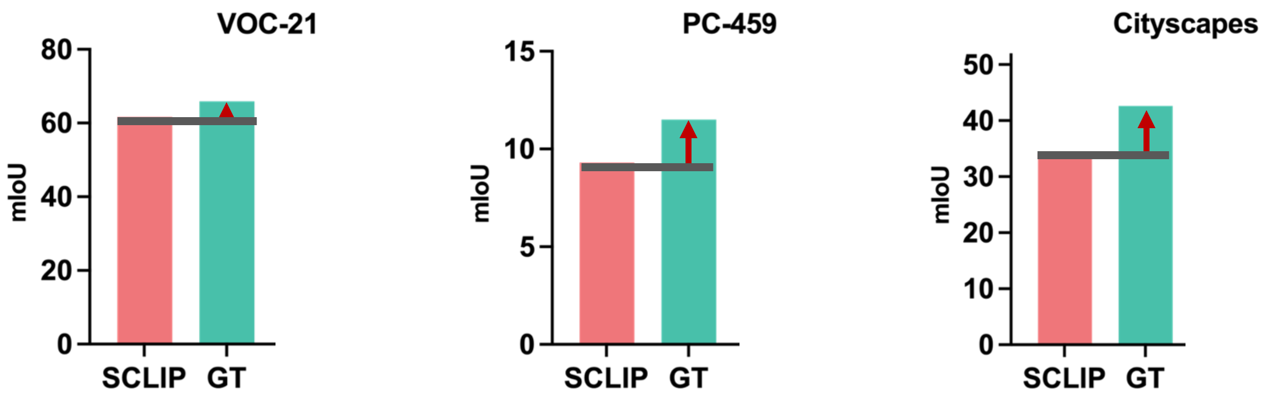
Why data quality matters?

- Comparing OVS performance between:
  - SCLIP, a representative method that modifies CLIP attention
  - Retrieving from GT segment-text of COCO Stuff with a simple strategy

Preliminary Study on Data Quality

Why data quality matters?

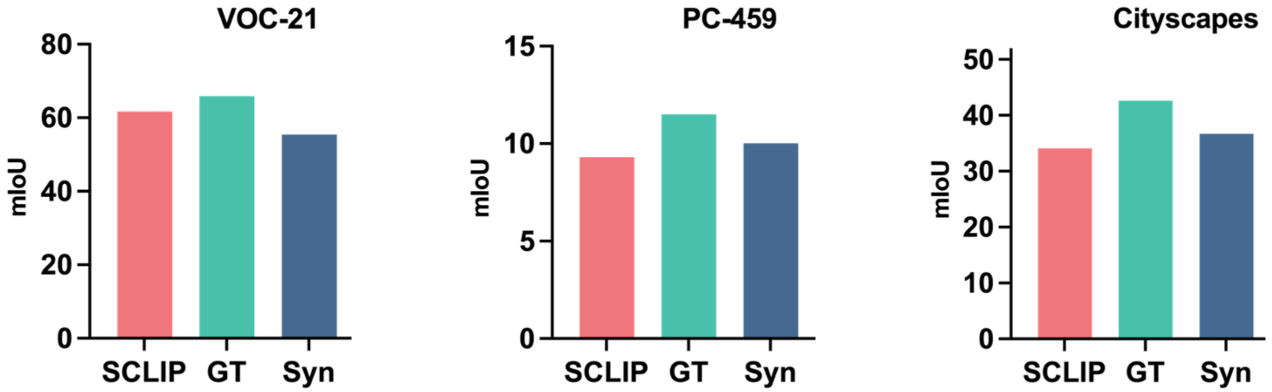
- Comparing OVS performance between:
  - SCLIP, a representative method that modifies CLIP attention
  - Retrieving from GT segment-text of COCO Stuff with a simple strategy



Preliminary Study on Data Quality

Why data quality matters?

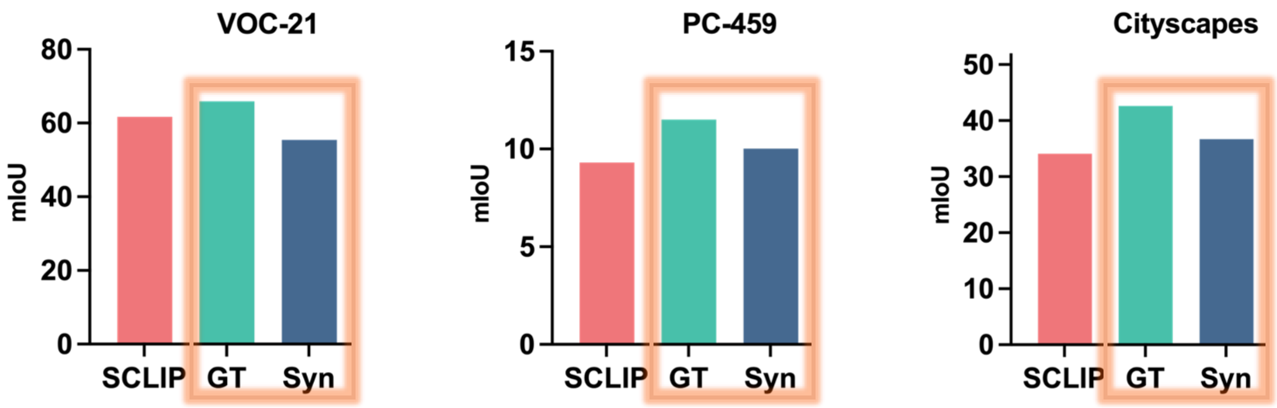
- Comparing OVS performance between:
  - SCLIP, a representative method that modifies CLIP attention
  - Retrieving from the synthetic (Syn) reference set of FreeDA



Preliminary Study on Data Quality

Why data quality matters?

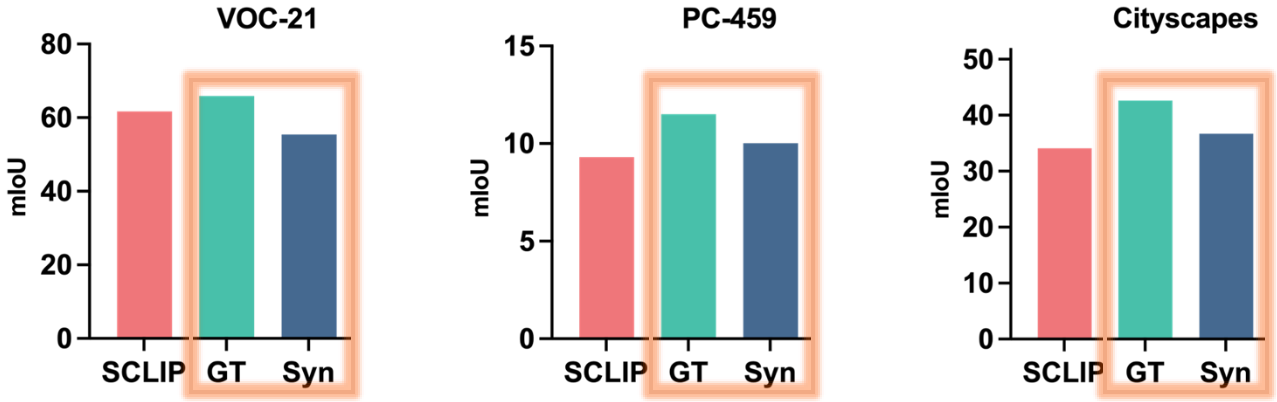
- Comparing OVS performance between:
  - SCLIP, a representative method that modifies CLIP attention
  - Retrieving from the synthetic (Syn) reference set of FreeDA



Preliminary Study on Data Quality

Why data quality matters?

- Comparing OVS performance between:
  - SCLIP, a representative method that modifies CLIP attention
  - Retrieving from the synthetic (Syn) reference set of FreeDA

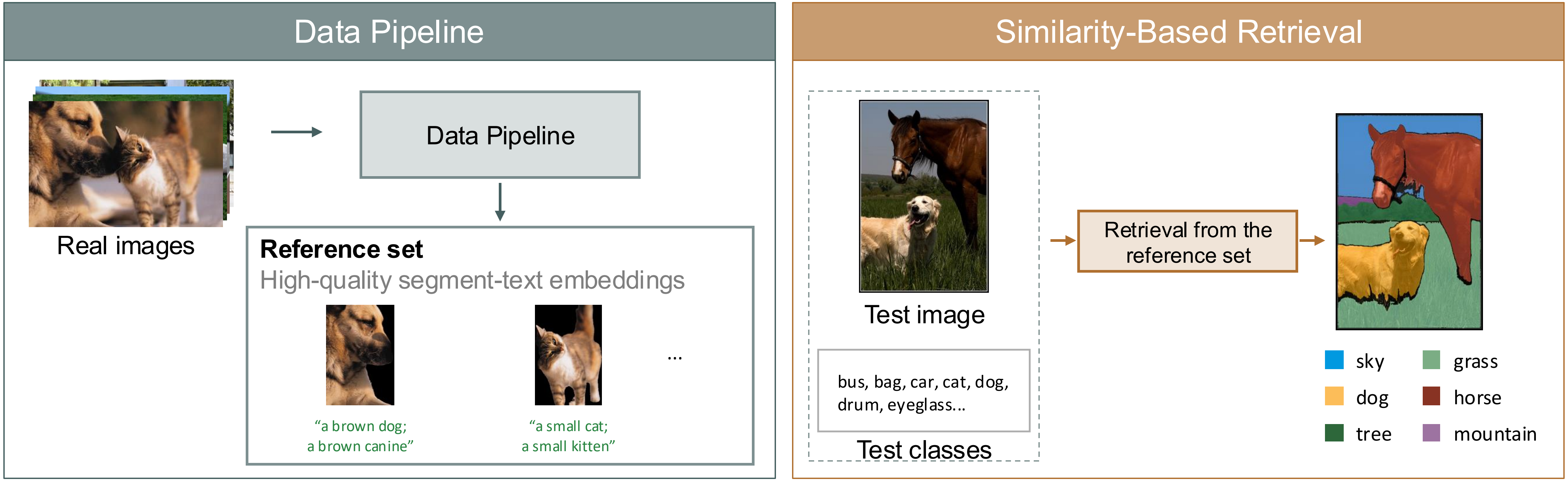


Without ground-truth annotations,

how to curate high-quality, densely annotated datasets?

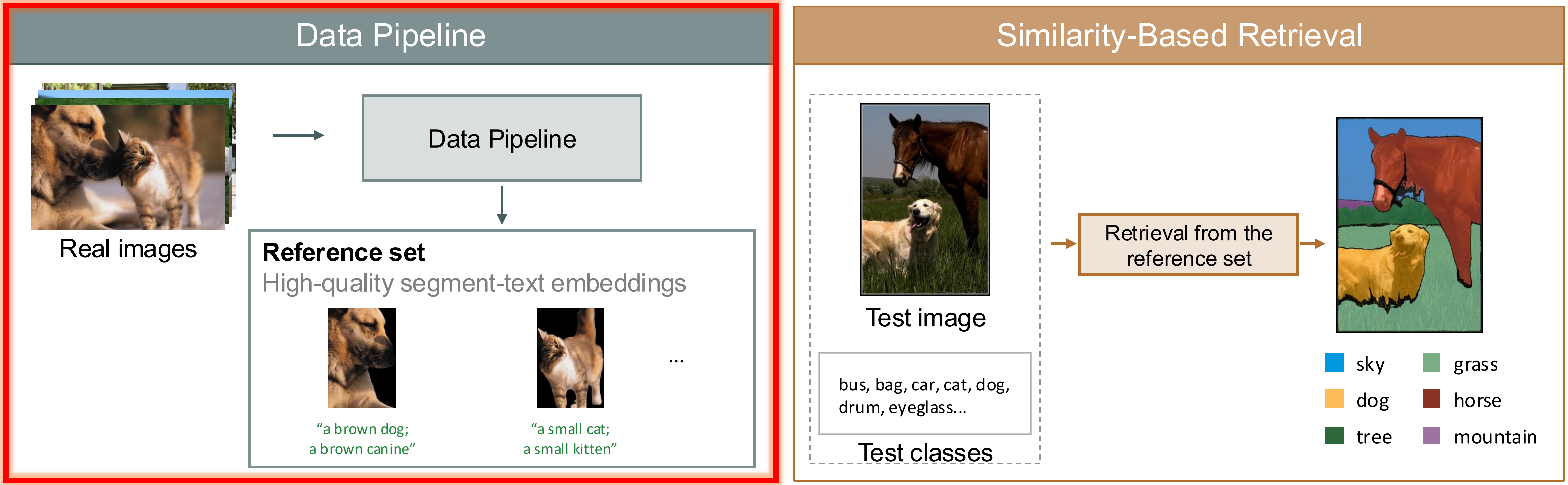
Our Approach

ReME: A Data-Centric Framework for Training-Free OVS



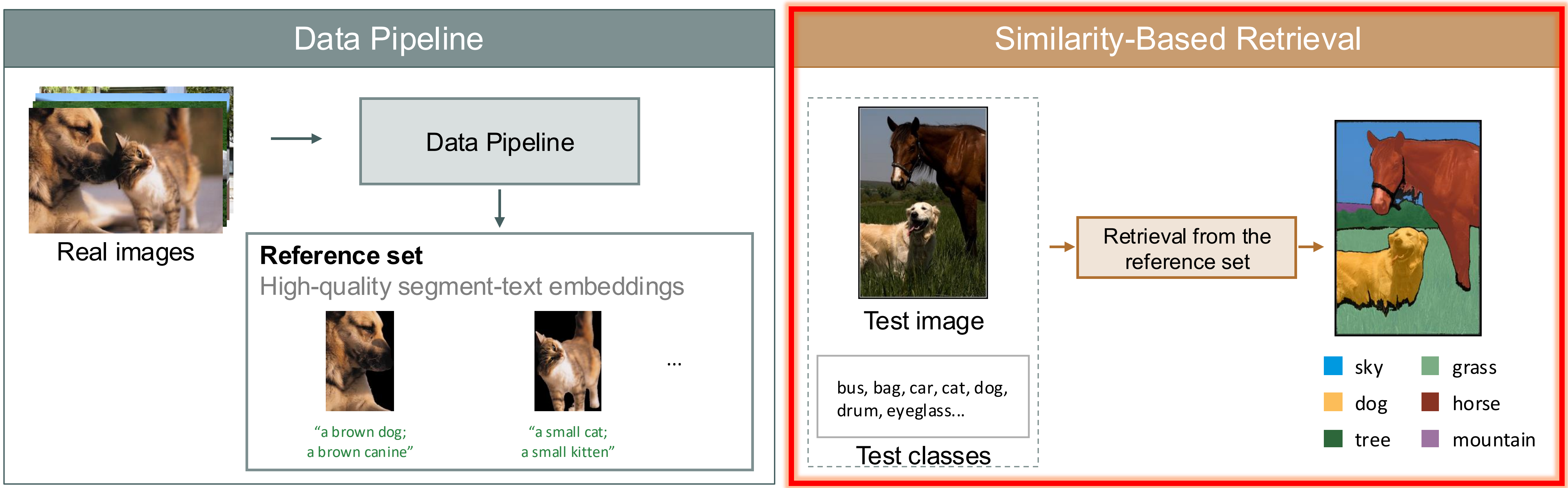
Our Approach

ReME: A Data-Centric Framework for Training-Free OVS



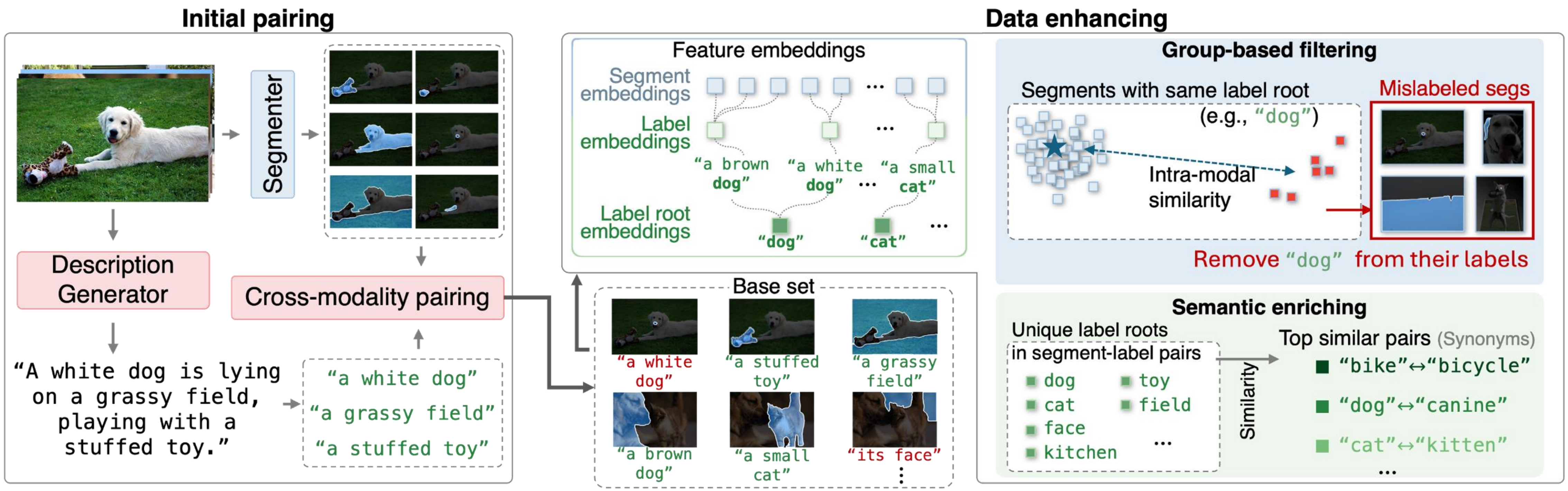
Our Approach

ReME: A Data-Centric Framework for Training-Free OVS



ReME Data Pipeline

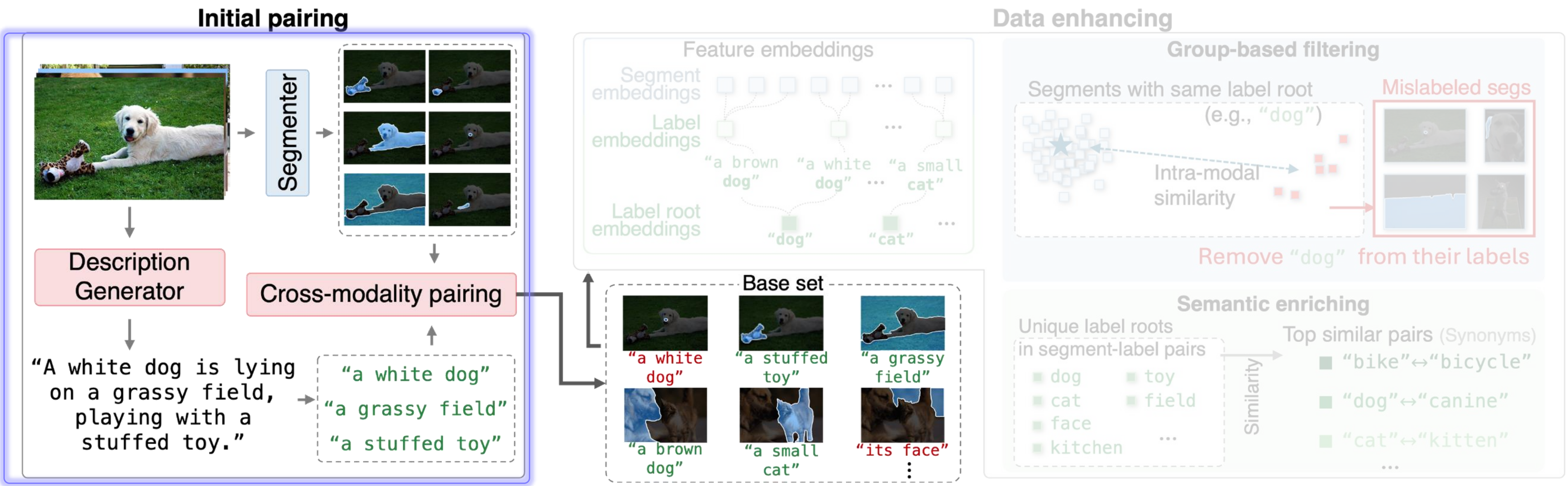
Goal:
Constructing a well-aligned, rich, and contextually relevant reference set with segment-text embeddings



ReME Data Pipeline

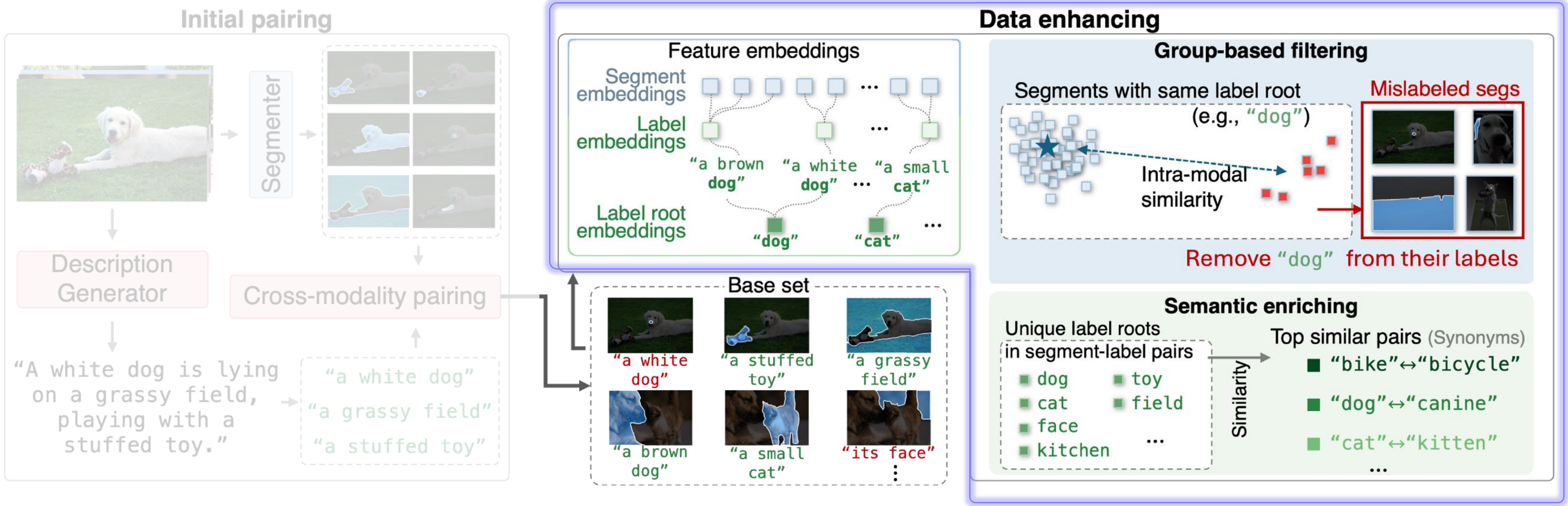
Initial pairing:

Obtain a diverse base set with segment-text pairs using images as input



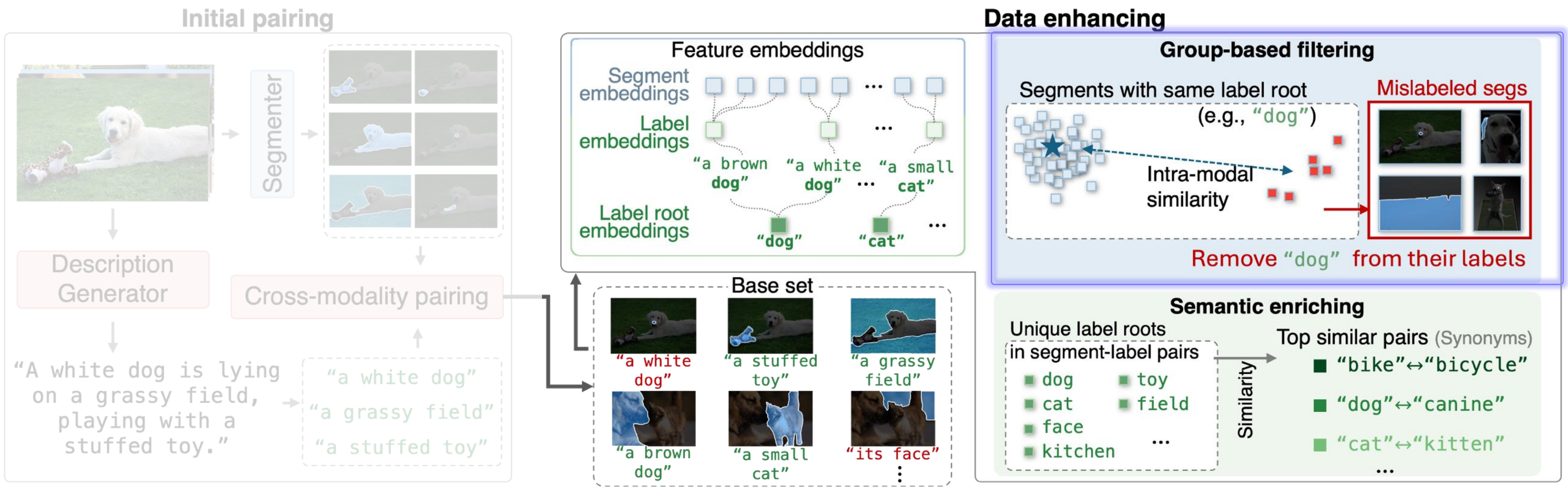
ReME Data Pipeline

Data enhancing:
Leverage the superior discriminativeness of intra-modal features to clean and enrich the reference set



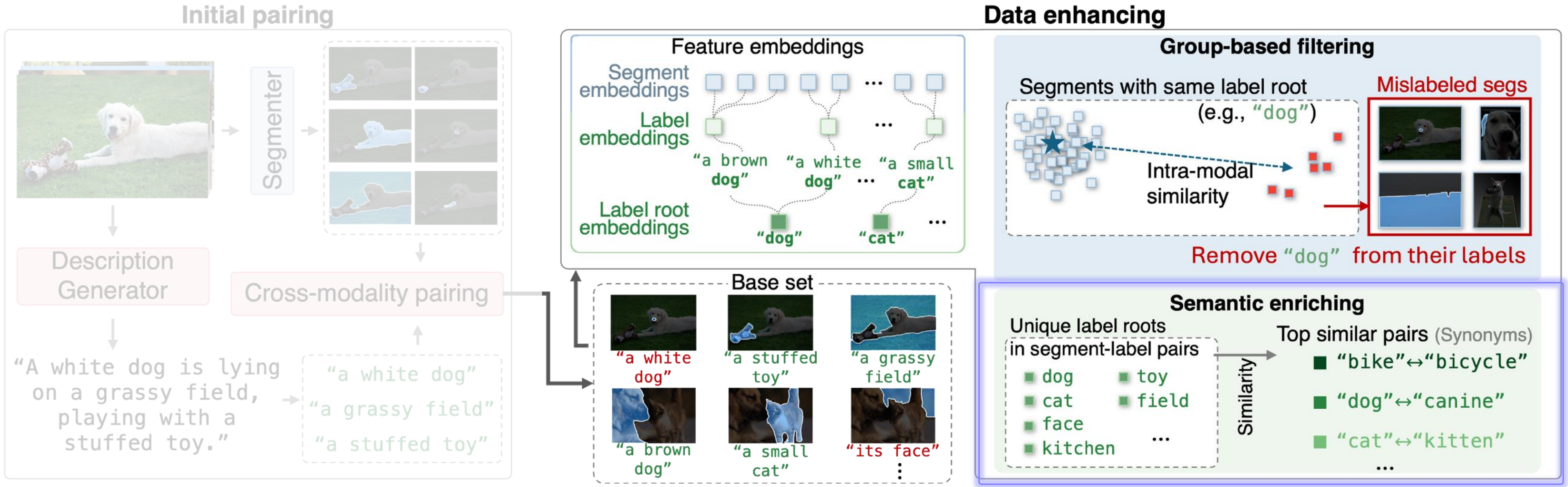
ReME Data Pipeline

Data enhancing:
Leverage the superior discriminativeness of intra-modal features to clean and enrich the reference set



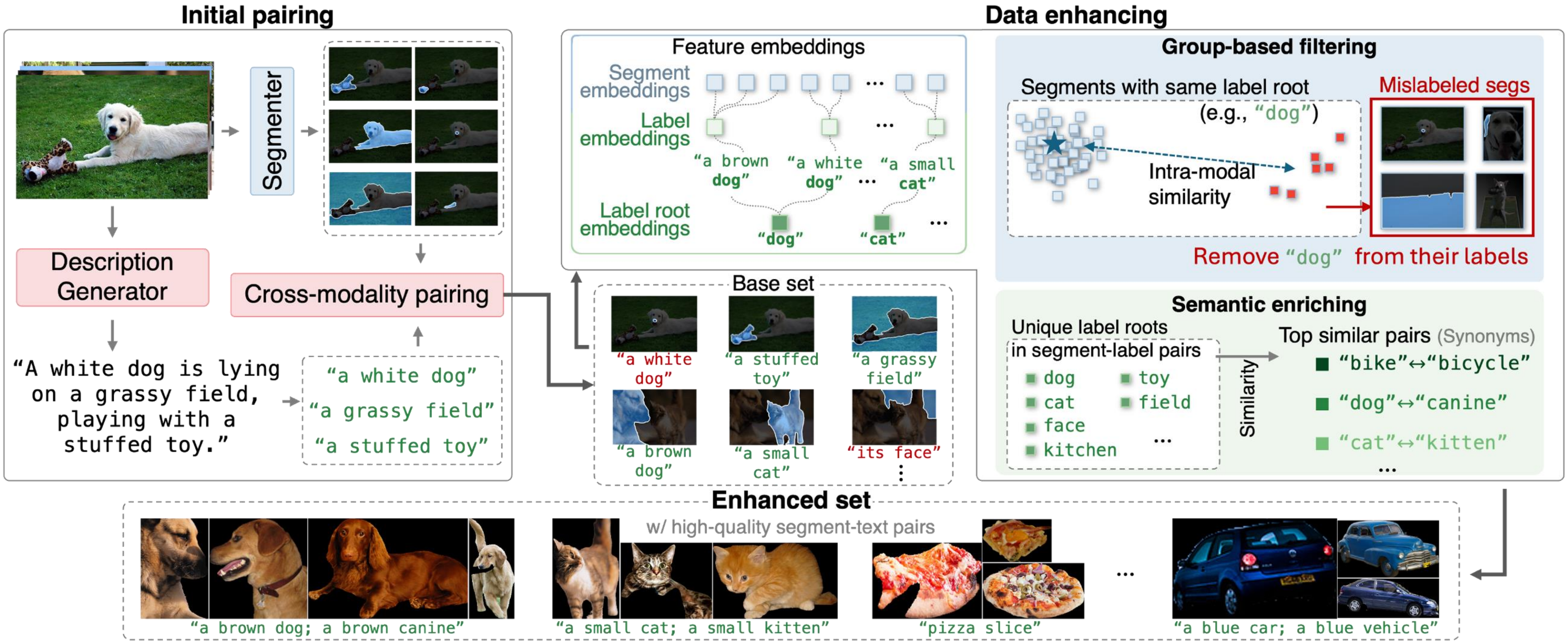
ReME Data Pipeline

Data enhancing:
Leverage the superior discriminativeness of intra-modal features to clean and enrich the reference set

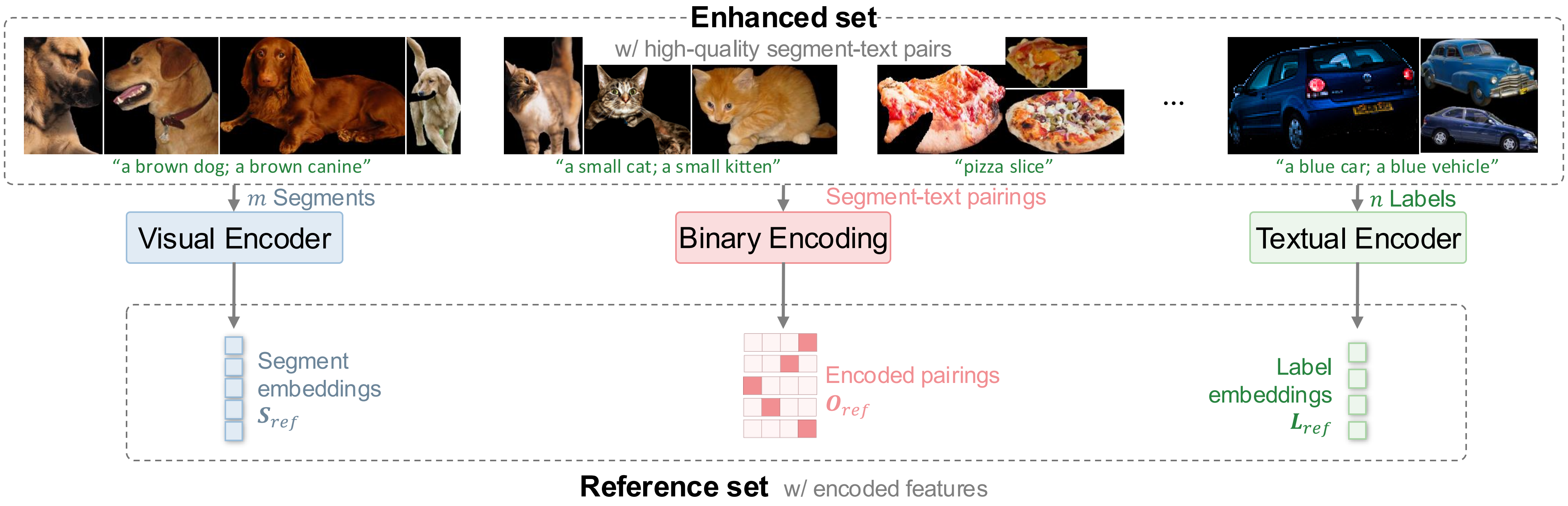


ReME Data Pipeline

Data enhancing:
Leverage the superior discriminativeness of intra-modal features to clean and enrich the reference set

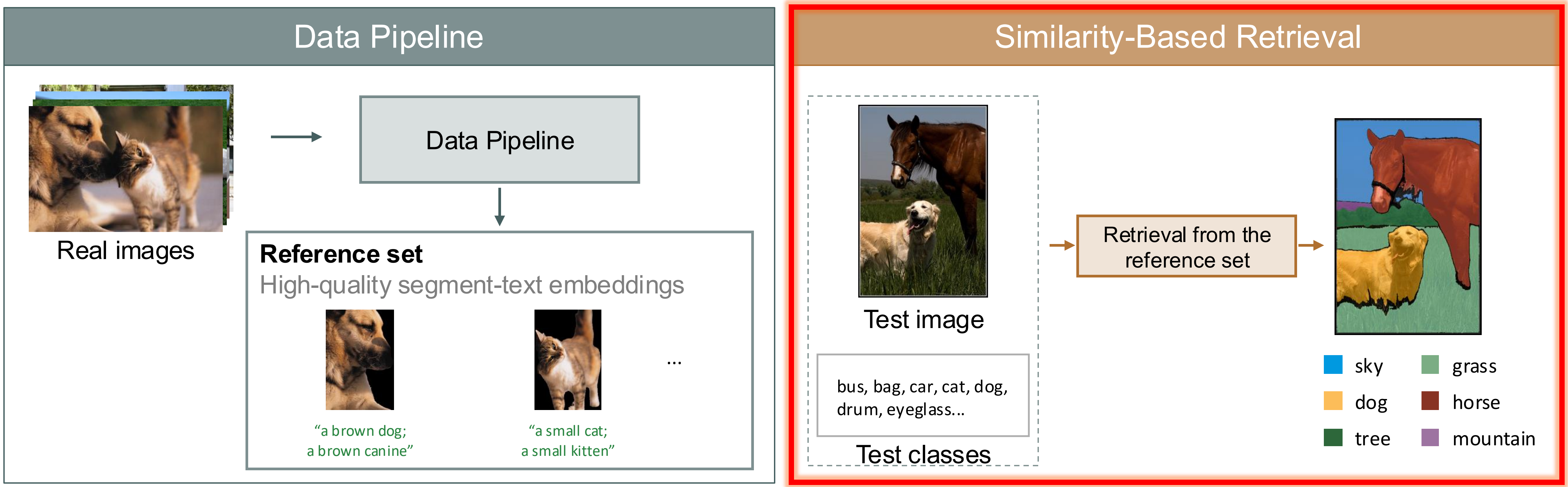


ReME Reference Set

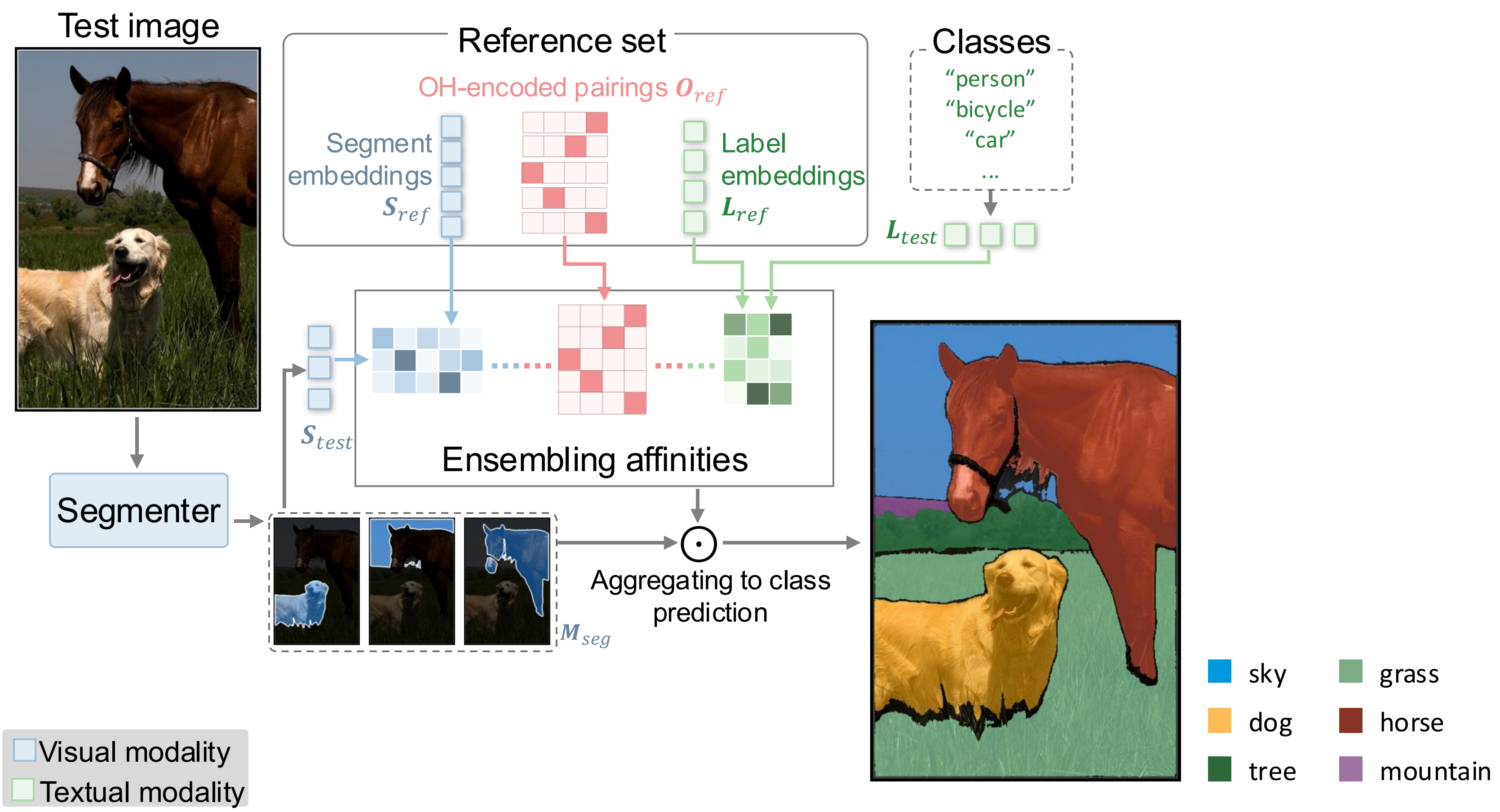


Our Approach

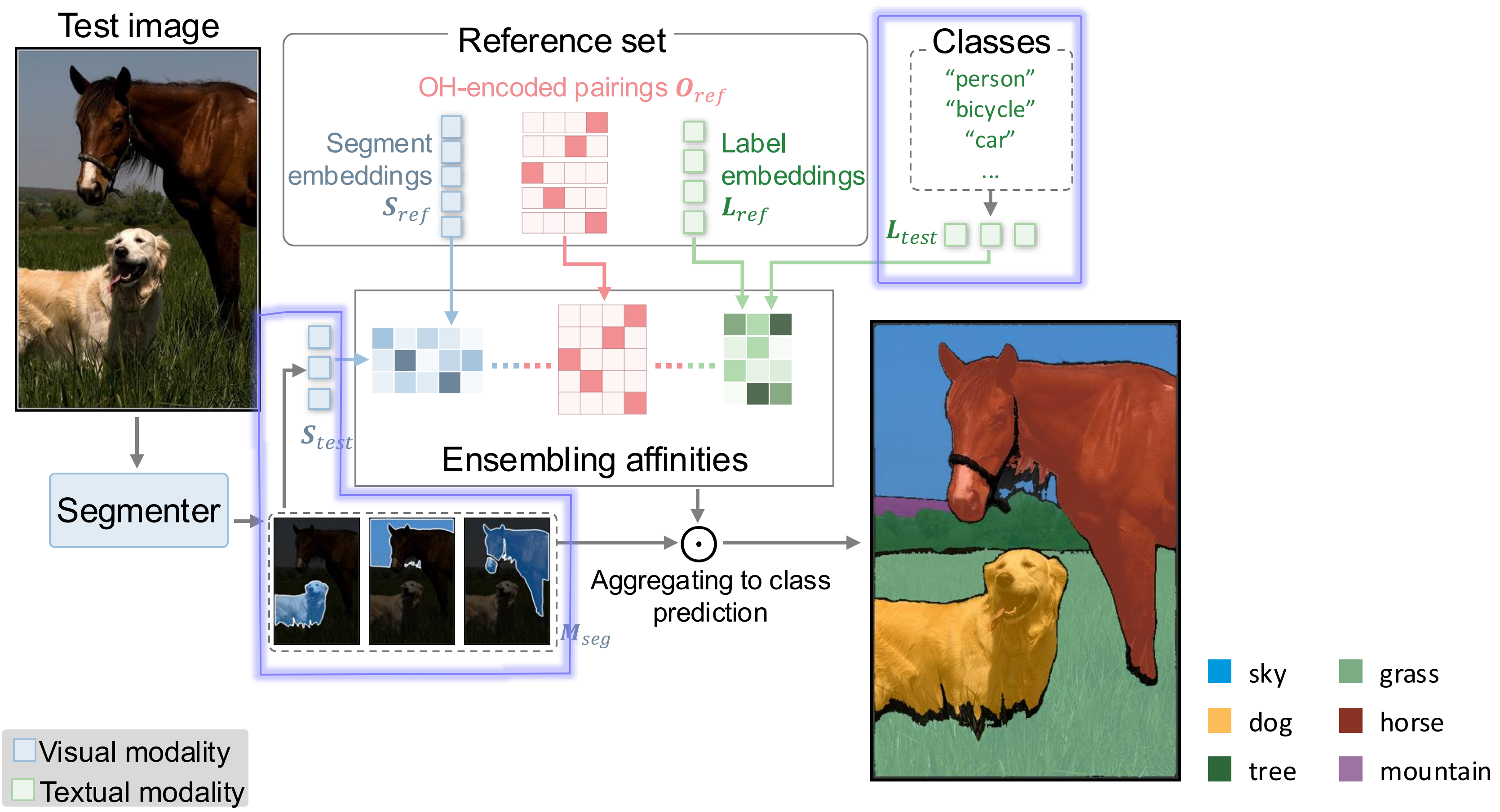
ReME: A Data-Centric Framework for Training-Free OVS



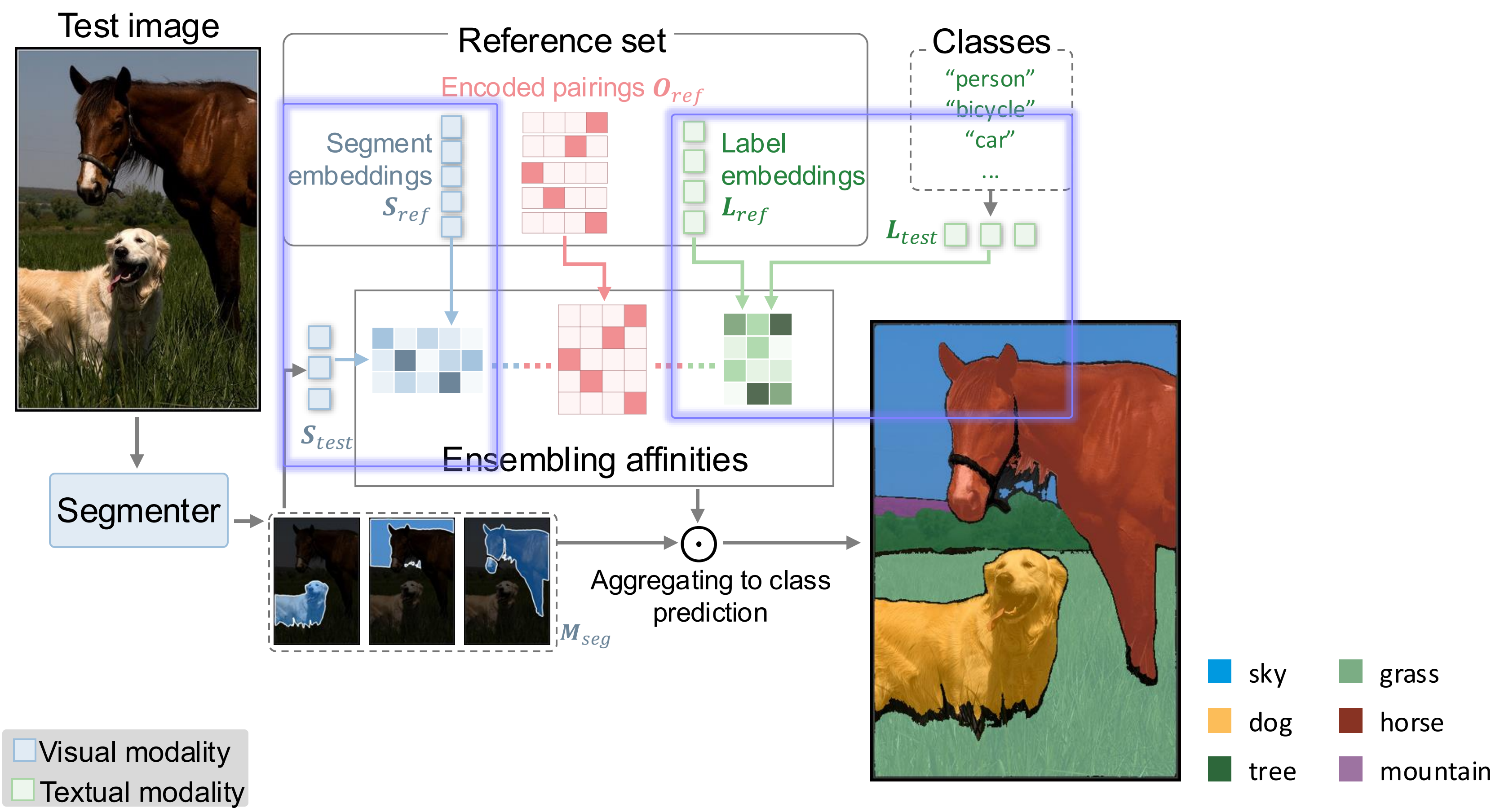
ReME Similarity-Based Retrieval



ReME Similarity-Based Retrieval

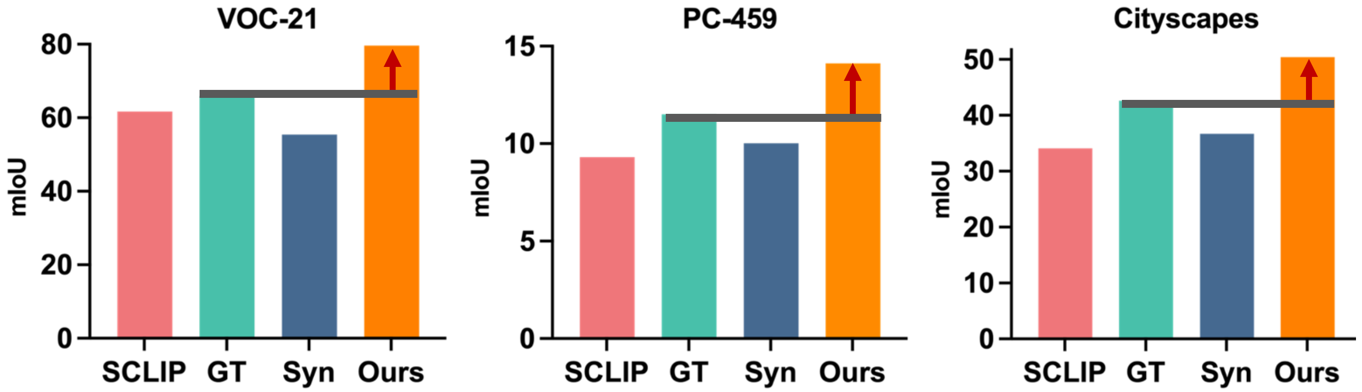


ReME Similarity-Based Retrieval



Data Quality Comparison

- Comparing OVS performance between:
  - Retrieving from GT segment-text of COCO Stuff with a simple strategy
  - Retrieving from the reference set of **ReME**



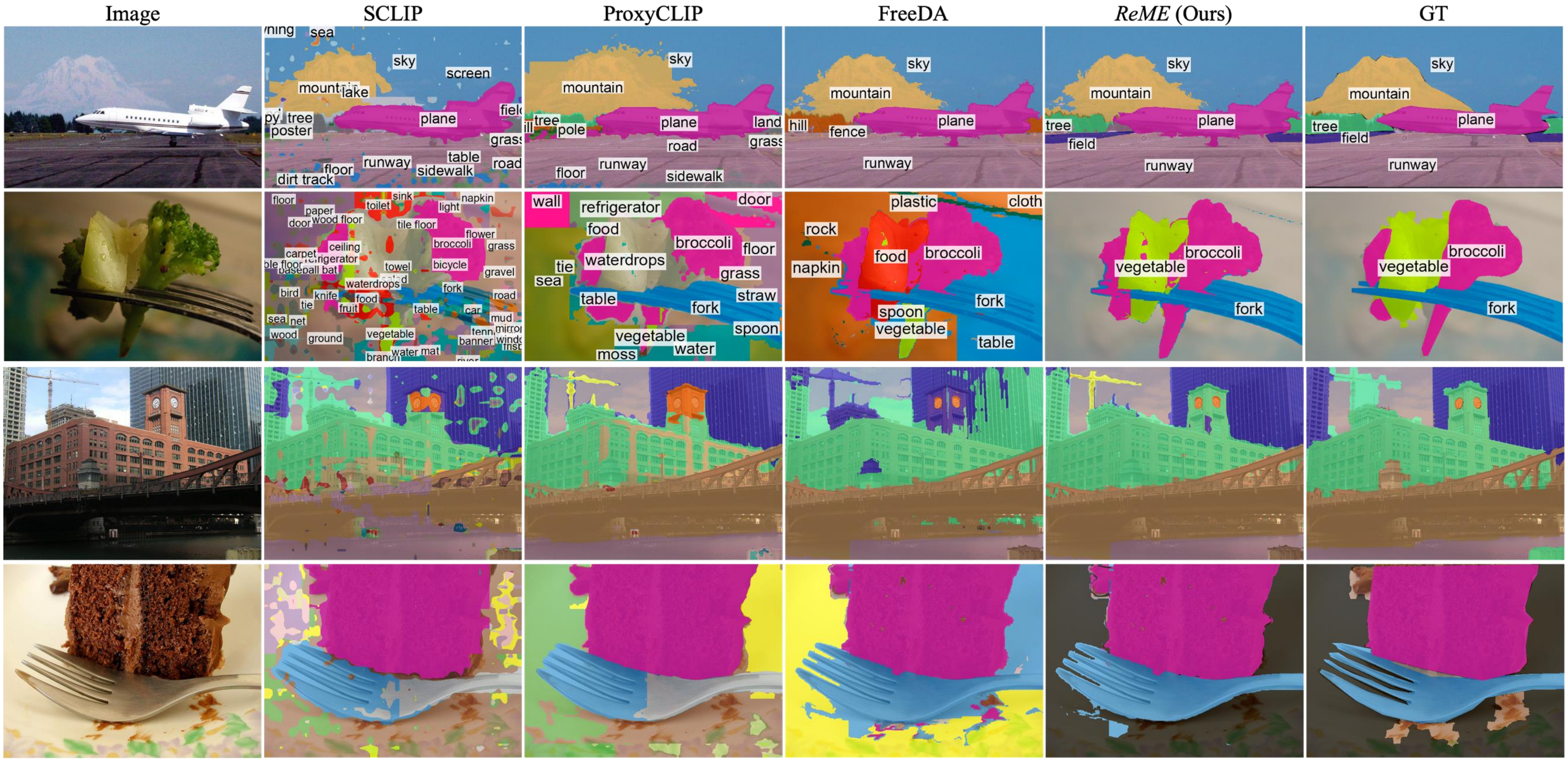
Our results even surpass retrieval from GT segment-text data

ReME Quantitative Results

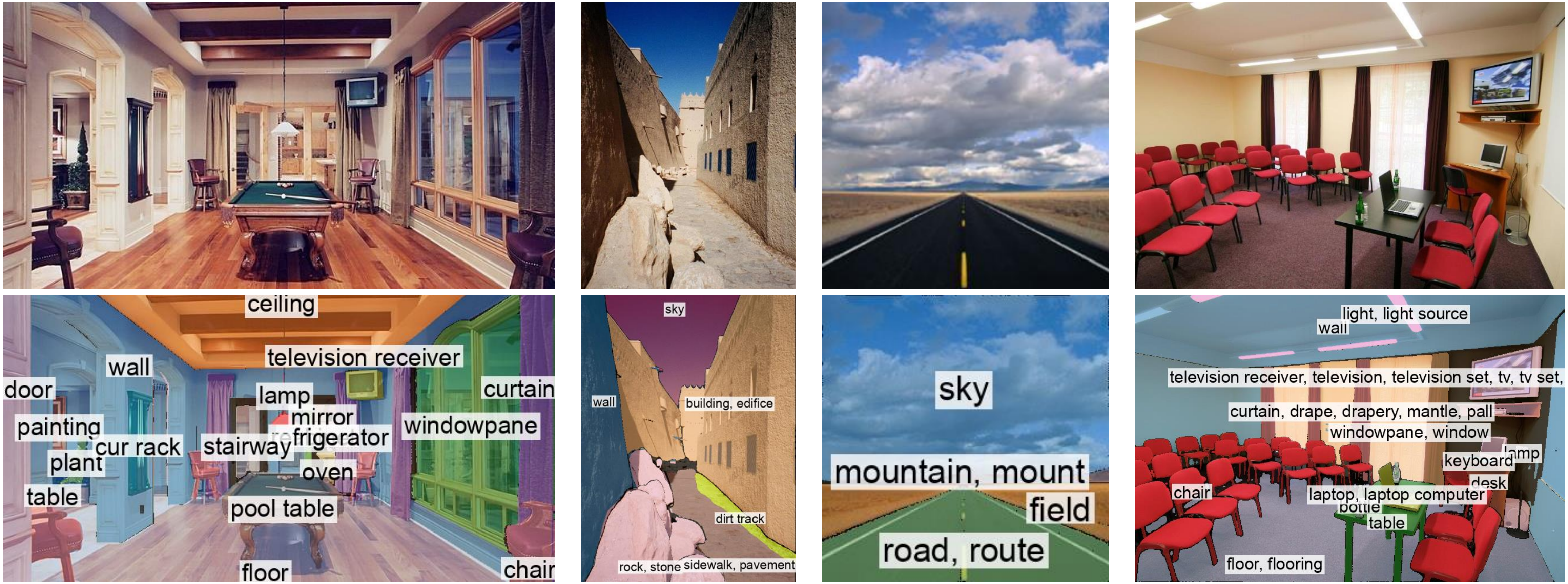
Methods	Post-processing	mIoU									
		VOC-20	VOC-21	City	PC-59	PC-60	Object	Stuff	A-150	PC-459	A-847
Training-free Methods without SAM											
GEM [4]	✗	46.2	24.7	-	32.6	21.2	-	15.1	10.1	4.6	3.7
MaskCLIP [73]	✓	74.9	38.8	12.6	25.5	23.6	20.6	14.6	9.8	-	-
ReCo [52]	✓	62.4	27.2	23.2	24.7	21.9	17.3	16.3	12.4	-	-
SCLIP [55]	✓	83.5	61.7	34.1	36.1	31.5	32.1	23.9	17.8	9.3	6.1
CaR [54]	✓	91.4	67.6	15.1	39.5	30.5	36.6	11.2	17.7	11.5	5.0
NACLIP [21]	✓	83.0	64.1	38.3	38.4	35.0	36.2	25.7	19.1	9.0	6.5
CLIPtrase [51]	✓	81.2	53.0	21.1	34.9	30.8	39.6	24.1	17.0	9.9	5.9
PnP [35]	✓	79.1	51.3	19.3	31.0	28.0	36.2	17.9	14.2	5.5	4.2
FreeDA [3]	✓	87.9	55.4	36.7	43.5	38.3	37.4	28.8	22.4	10.2	5.3
ProxyCLIP [27]	✗	83.2	60.6	40.1	37.7	34.5	39.2	25.6	22.6	11.2	6.7
DiffSegmenter [57]	✓	71.4	60.1	-	27.5	25.1	37.9	-	-	-	-
OVDiff [23]	✓	80.9	68.4	23.4	32.9	31.2	36.2	20.3	14.1	12.0	6.6
***ReME* (Ours)**	✗	**92.3**	**79.6**	**50.4**	**44.9**	**41.6**	**45.5**	**33.1**	**26.1**	**14.1**	**8.4**
***ReME* (Ours - VOC)**	✗	84.7	**75.0**	**43.9**	40.9	**38.7**	**40.8**	22.6	**25.2**	12.8	**8.3**
***ReME* (Ours - ADE)**	✗	84.3	**72.3**	**42.1**	**44.0**	**39.7**	35.8	27.0	**26.0**	13.2	**8.6**
Training-free Methods with SAM											
RIM [59]	✗	77.8	-	-	34.3	-	44.5	-	-	-	-
CaR w/ SAM [54]	✗	-	70.2	16.9	40.5	31.1	37.6	12.4	17.9	11.8	5.7
CLIPtrase w/ SAM [51]	✗	82.3	57.1	-	36.4	32.0	44.2	24.8	17.2	10.6	6.0
ProxyCLIP w/ SAM [27]	✗	80.4	59.3	37.0	37.0	33.6	35.4	25.0	19.1	6.9	4.8
CorrCLIP [70]	✗	91.6	74.1	47.7	45.5	40.3	43.6	30.6	-	-	-
***ReME* w/ SAM (Ours)**	✗	**93.2**	**82.2**	**59.0**	**53.1**	**44.6**	**48.2**	**33.3**	**28.2**	**15.8**	**8.8**

ReME outperforms all training-free baselines across ten benchmark datasets

ReME Qualitative Results



ReME Qualitative Results



ReME Qualitative Results

In-the-wild results obtained by prompting **ReME** with diverse free-form textual inputs.

Portable computer
A cute tabby cat
A laying person
Cozy brown couch

Energetic golden retriever in motion
Purple agility tunnel

Domestic cattle
A static cattle egret

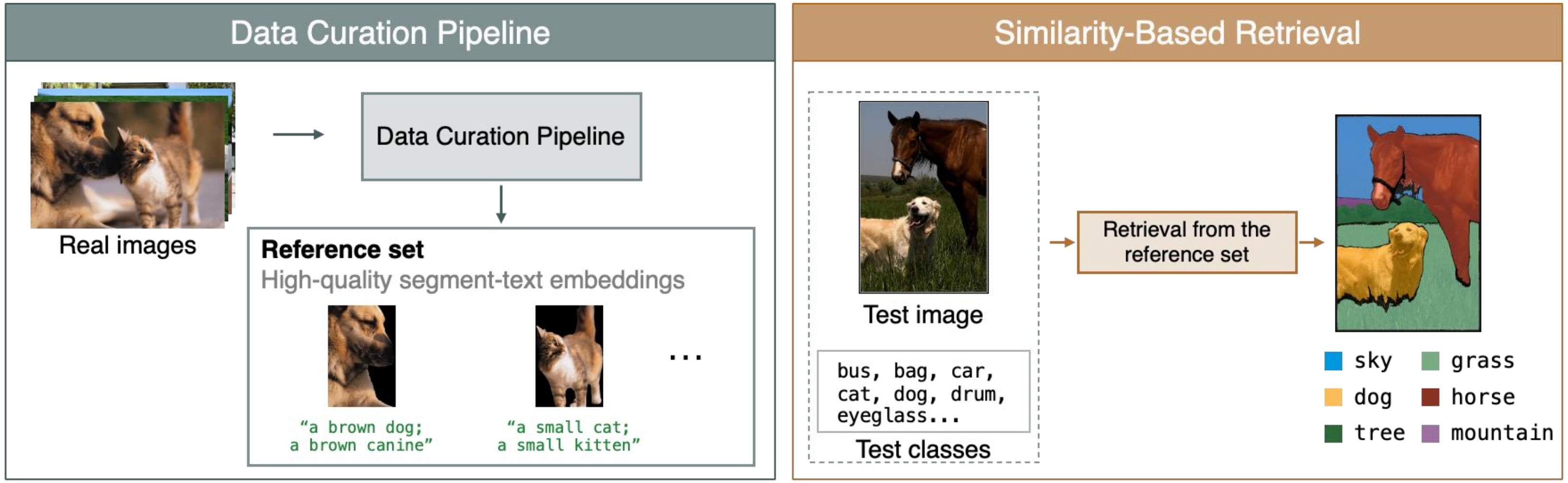
A flying propeller aircraft

Sleeping lions
Huge boulders

Sunshade
Beach lounger
Peaceful blue ocean
Towel

Conclusion

A Data-Centric Framework for Training-Free Open-Vocabulary Segmentation



+ Training-Free, Flexible, Data-Centric OVS Framework

+ Scalable Data Pipeline Providing High-Quality Segment-Text Embeddings w/o Human Annotations

+ Open-Vocabulary Segmentation Surpasses all Training-Free Baselines across Ten Benchmark Datasets

SCAN ME

- **Thanks!**
- **Please check out the full paper for more details!**

UCDAVIS

VISUALIZATION & INTERFACE DESIGN INNOVATION