CVPR · Denver 2026

CVPR 2026 · Open-Vocabulary Panoptic Segmentation

Mitigating Objectness Bias and Region-to-Text Misalignment

OVRCOAT

a simpler path to open-vocabulary panoptic segmentation

Nikolay Kormushev[1,2] **Josip Šarić**[1,3] **Matej Kristan**[1]

[1] University of Ljubljana, FRI · [2] ETH Zürich · [3] University of Zagreb, FER

EMAI Erasmus Mundus joint Master in Artificial Intelligence

FRI UNIVERSITY OF LJUBLJANA Faculty of Computer and Information Science

MOTIVATION

Closed vocabularies don't survive the real world

Panoptic segmentation = pixel labels + instance separation.

Conventional models = fixed vocabulary.

→ Real world needs **open-vocabulary** recognition.

SOTA performance on Open-Vocabulary Panoptic Segmentation

MAFT+ [ECCV'24] · FC-CLIP [NeurIPS'23]

The open-vocabulary recipe

1. Generate mask proposals
2. Reject low-objectness masks
3. Classify the rest using a VLM (e.g. CLIP)

Two bottlenecks

① Mask selection bias

Objectness heads trained on **closed vocabularies**.

Trained on COCO → **“painting” in images not labels** → **Predicted as background**

Šarić et al., What Holds Back OV Segmentation?, ICCVW 2025

② Region-to-text misalignment

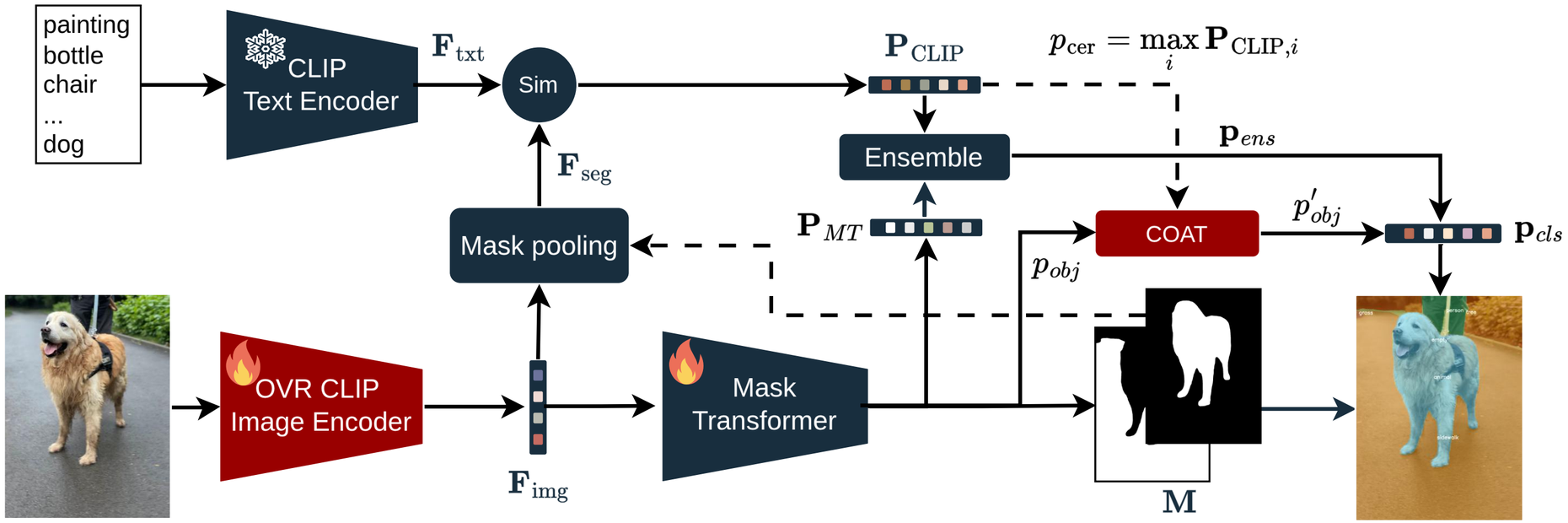
VLMs trained on image–text pairs.
→ Label embeddings ***not aligned*** with regions.

CLIP [Radford et al., ICML'21] · ALIGN [Jia et al., ICML'21]

SOTA performance on Open-Vocabulary Panoptic Segmentation

MAFT+ [ECCV'24] · FC-CLIP [NeurIPS'23]

OVRCOAT = OVR (training) + COAT (inference)



Built on Mask2Former [CVPR'22] + ConvNeXt-L CLIP [LAION-2B]

OVR — open-vocabulary mask-to-text refinement *(fine-tunes CLIP for region understanding)*

COAT — CLIP-conditioned objectness adjustment *(test-time bias correction, no extra training)*

COAT — using CLIP to debias objectness

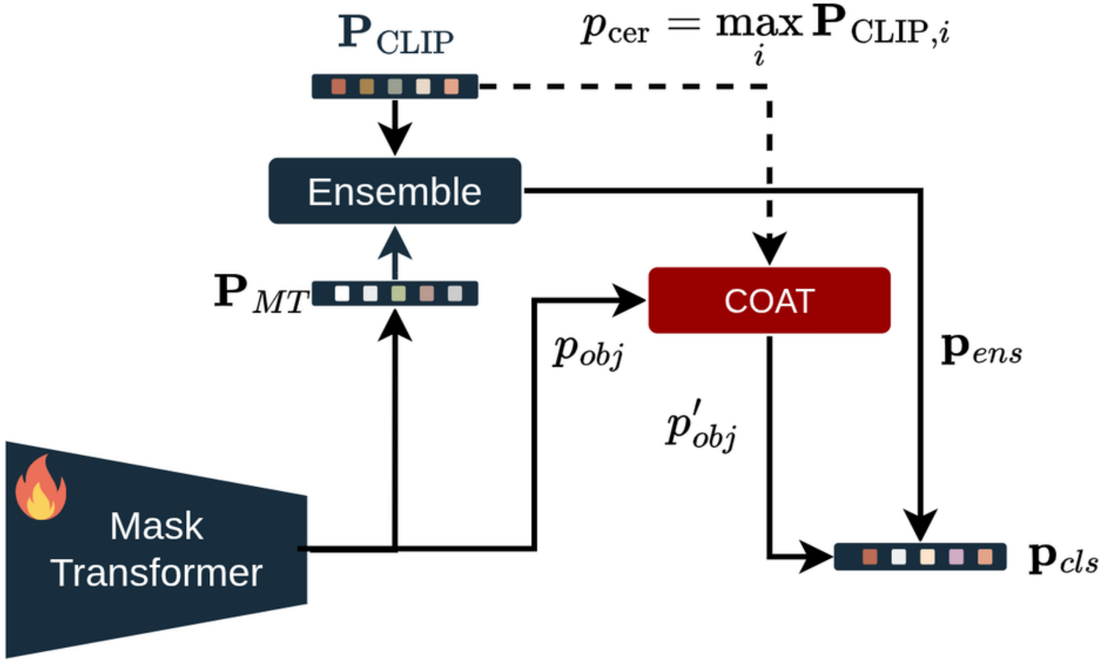
Insight. Void token suppresses unseen categories.

Fix. Leverage CLIP web-scale pretraing to update confidence.

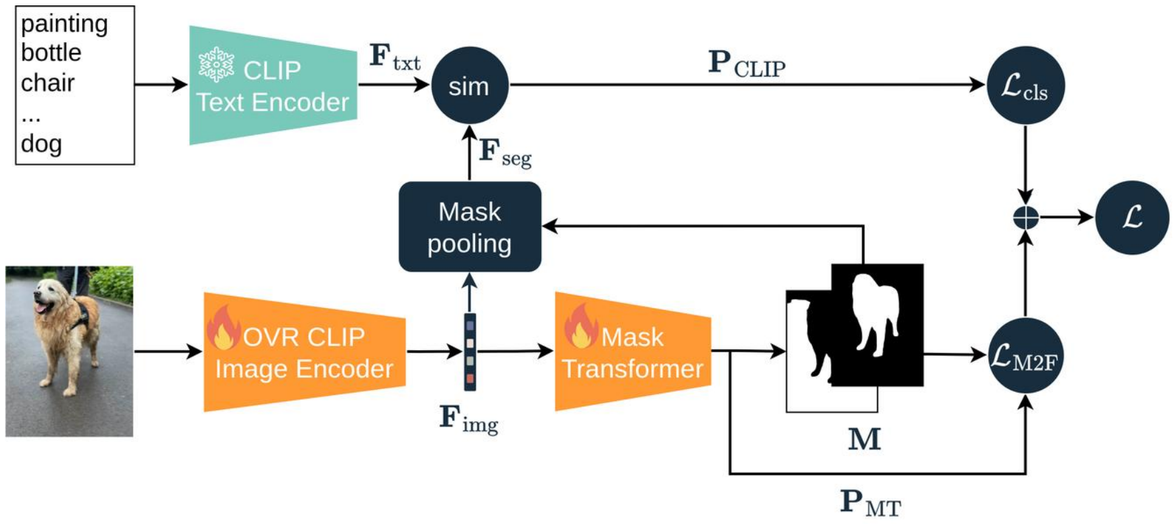
Effect. Rescues masks the transformer would reject.

$$p'_{obj} = 1 - (1 - \gamma \cdot p_{cer})(1 - p_{obj})$$

p_{cer} = max CLIP class confidence · $\gamma = 0.5$ (CLIP trust factor)



OVR — refining CLIP for region-level alignment



A combined objective

$$L = \alpha \cdot L_{cls} + L_{M2F}$$

Two-stage training protocol

1 **Frozen CLIP encoder**
Pre-train mask generation without disrupting CLIP's embedding space.

2 **Unfrozen CLIP (final MLP + norm frozen)**
Jointly tunes mask-text alignment and proposal accuracy.

→ **Stronger region-level alignment, improves classification.**

RESULTS

New SOTA on three OOV benchmarks

ADE20K

28.6

PQ

+5.5%

Mapillary Vistas

19.6

PQ

+7.1%

Cityscapes

45.3

PQ

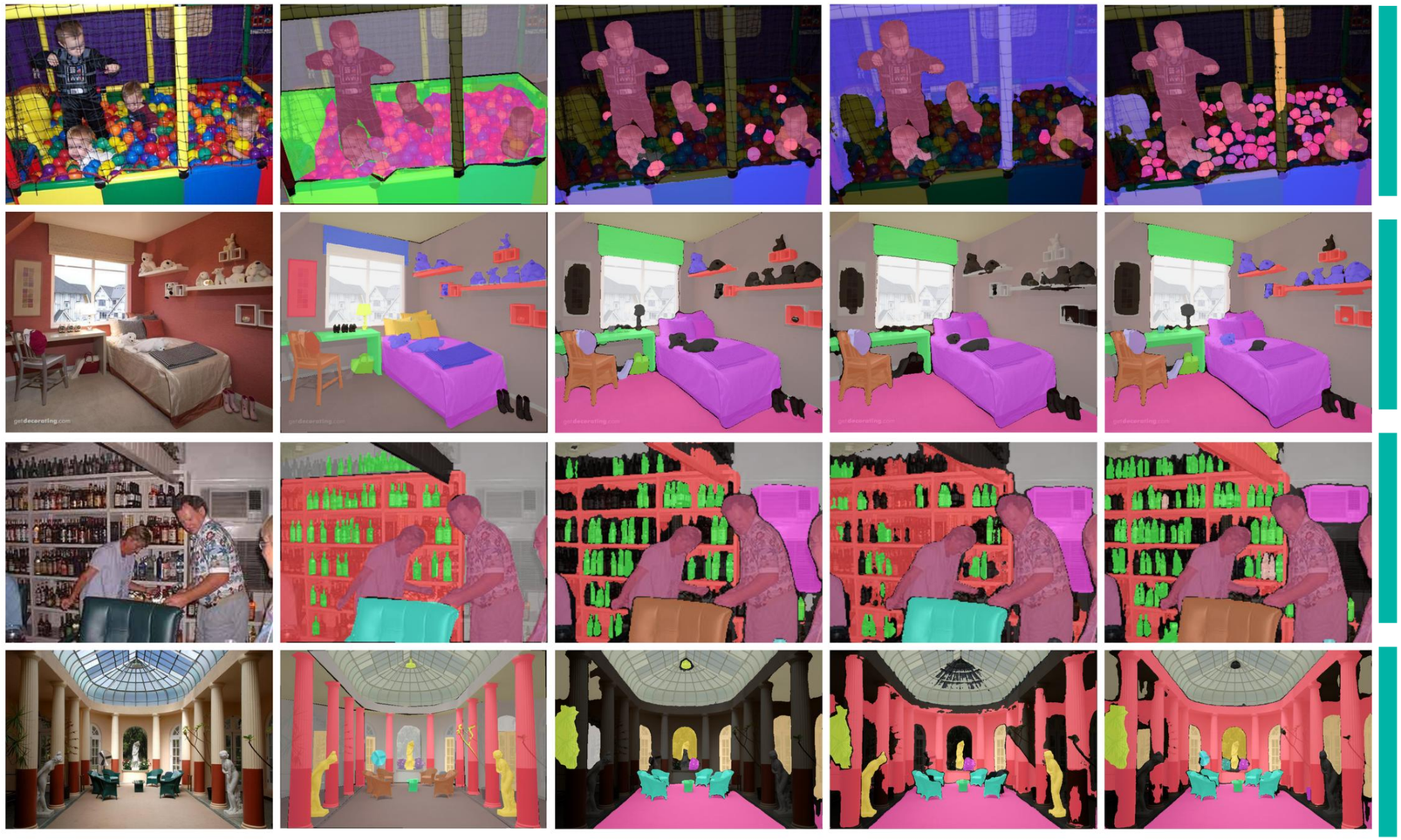
+3.0%

Method	ADE20K	Mapillary	Cityscapes	COCO
ODISE [CVPR'23]	23.4	14.2	23.9	55.4
FC-CLIP [NeurIPS'23]	26.8	18.3	44.0	54.4
MAFT+pan [ECCV'24]	27.1	15.7	38.3	50.3
OVRCOAT (ours)	**28.6**	**19.6**	**45.3**	**54.6**

PQ on out-of-vocabulary benchmarks. COCO is in-vocabulary (training set).

Recovering masks the SOTA misses

no mask	armchair
awning	ball
bar	basket
bed	blanket
bottle	box
ceiling	ceiling fan
chair	coffee table
column	cup
cushion	desk
door	floor
flower	fountain
lamp	microwave
painting	person
plant	plant pots
plaything	pole
rug	sculpture
seat	shelf
towel	wall
window screen	windowpane



Input | Ground truth | FC-CLIP | MAFT+ | OVRCOAT (ours)

Small objects

Toy balls in the bedroom.

Missed instances

Bags & toys recovered.

Fine objects

Individual bottles, not shelf.

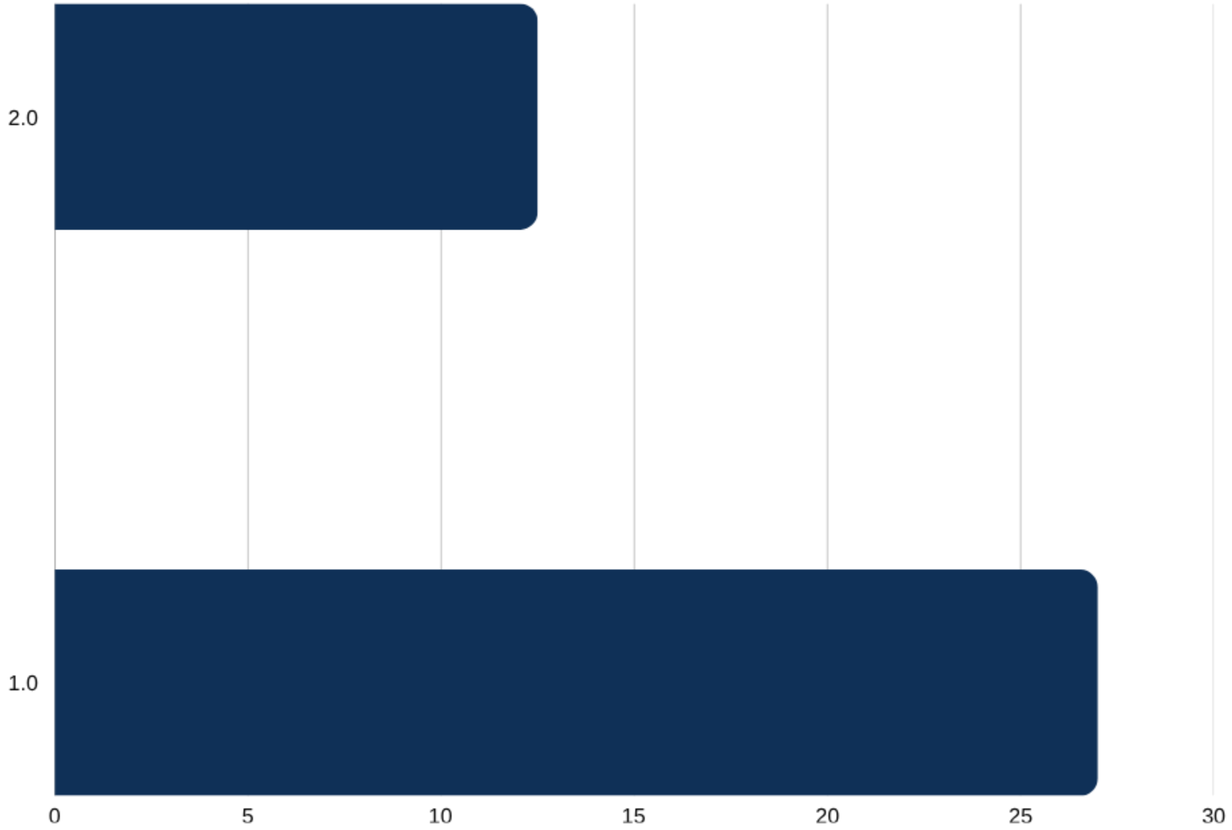
Architecture

Doors & pillars in the museum.

EMAI Erasmus Mundus joint Master in Artificial Intelligence
FRI UNIVERSITY OF LJUBLJANA Faculty of Computer and Information Science

Less memory, more PQ

Training memory (GB / image)



→ *~56 % less memory than MAFT+[ECCV'24] (previous SOTA)*

Component ablation (PQ)

Variant	ADE20K	Mapillary	Cityscapes
FC-CLIP baseline	26.8	18.3	44.0
+ COAT only	27.6	18.8	44.6
+ OVR only	27.6	19.2	44.5
OVRCOAT (both)	**28.6**	**19.6**	**45.3**

Seen vs Unseen on ADE20K

Seen classes: **–0.05 pp**
Unseen classes: **+3.9 pp (+25 % rel.)**
'Painting' class: **+192 % rel. PQ**

TAKEAWAYS

Two simple, modular fixes - a cleaner path to open-vocabulary panoptic perception

Simple beats complex

A simplified architecture surpasses far more involved recent approaches.

Better foundation

A simpler architecture opens space for new modules and faster development iterations.

Practical

~56% less training memory than MAFT+ (12.5 vs 27.0 GB / image).

BACKUP

References

[1] **Cheng et al..** Masked-attention Mask Transformer for Universal Image Segmentation *CVPR 2022*

[2] **Radford et al..** CLIP: Learning Transferable Visual Models From Natural Language Supervision *ICML 2021*

[3] **Jia et al..** ALIGN: Scaling Up Visual and Vision-Language Representation Learning *ICML 2021*

[4] **Liu et al..** A ConvNet for the 2020s (ConvNeXt) *CVPR 2022*

[5] **Kirillov et al..** Panoptic Segmentation *CVPR 2019*

[6] **Xu et al..** ODISE: OV Panoptic Segmentation with Diffusion Models *CVPR 2023*

[7] **Yu et al..** FC-CLIP: Open-Vocabulary Segmentation with Single Frozen Convolutional CLIP *NeurIPS 2023*

[8] **Jiao et al..** MAFT+: Collaborative Vision-Text Representation Optimizing for OV Seg. *ECCV 2024*

[9] **Šarić et al..** What Holds Back Open-Vocabulary Segmentation? *ICCVW 2025*

[10] **Zhou et al..** Scene Parsing through ADE20K Dataset *CVPR 2017*

[11] **Cordts et al..** The Cityscapes Dataset for Semantic Urban Scene Understanding *CVPR 2016*

[12] **Neuhold et al..** The Mapillary Vistas Dataset *ICCV 2017*

[13] **Lin et al..** Microsoft COCO: Common Objects in Context *ECCV 2014*

Full reference list in the paper.