Metric-Guided Feature Fusion of Visual Foundation Models for Segmentation Tasks

Supplementary Material

A. Metric Details

A.1. Hyperparameter Sensitivity

Our metrics involve several hyperparameters. To verify that the structure–edge characterization is robust to hyperparameters choices, we vary each within a range while keeping others fixed: SFC grid size $K \in \{8, 12, 16, 20\}$; SCS PCA dimensions $\in \{16, 32, 64\}$ and cluster counts $k \in \{4, 6, 8, 10, 12\}$; EC/NC radii $r_{\text{in}} \in \{2, 3, 4\}$, $r_{\text{out}} \in \{6, 7, 8\}$; FC cutoff $\rho_{\text{low}} \in \{0.10, 0.15, 0.20\}$; SP threshold $\tau \in \{0.4, 0.5, 0.6\}$. In these cases, absolute metric values shift but the diagnosed structure–edge bias of each encoder remains consistent, confirming that the characterization is stable across reasonable hyperparameter choices.

A.2. Cross-dataset consistency of SC/EF profiles.

In the main paper, Tab. 5 presents the SC and EF profiles of several VFMs on Cityscapes. The COCO results reported in Tab. 9 exhibit a highly consistent pattern across datasets. All encoders preserve their bias on both the SC and EF axes, despite the substantial domain shift between COCO and Cityscapes. This indicates that the structure–edge characterization revealed by our assessment metrics reflects intrinsic properties of the pretrained VFM encoders rather than dataset-specific artifacts. Such cross-dataset stability further supports using these metrics to guide encoder pairing and OS selection in our fusion framework.

Table 9. SC/EF profiles of frozen VFM encoders across output strides on COCO. Higher SC indicates stronger semantic coherence; higher EF indicates stronger boundary sensitivity.

Metric	SC				EF			
OS	4	8	16	32	4	8	16	32
DINOv2	0.66	0.62	0.55	0.47	1.95	3.37	5.30	8.09
DINOv3	0.75	0.74	0.69	0.61	0.92	1.19	2.22	3.54
SAM	0.73	0.70	0.62	0.53	2.62	3.15	4.27	5.68
SAM2	0.61	0.51	0.42	0.52	3.57	5.90	12.83	5.97
Swin-B	0.74	0.71	0.68	0.64	1.25	2.58	1.74	2.17

B. Additional Experimental Analysis

B.1. Per-class Analysis of Injection Stride

Tab. 10 provides per-class AP on Cityscapes when injecting a single frozen SAM2 feature stage into a fine-tuned DINOv2 or DINOv3 master backbone, complementing the summary in the main paper (Tab. 6). For both master encoders, injection at OS=16 yields the best average AP (38.7 for DINOv2, 39.1 for DINOv3), with gains most pronounced on boundary-sensitive classes such as person (+9.7/+4.0), bicycle (+4.4/+1.6), and motorcycle (+2.2/+1.3). Large rigid categories also benefit (e.g., car +7.4/+2.6), while injection at other strides produces smaller or inconsistent improvements. This confirms that the benefit is stage-specific and aligns with SAM2's EF score (Tab. 5), which peaks at OS=16.

Table 10. Per-class AP on Cityscapes with frozen SAM2 as auxiliary. Each column injects SAM2 features at a single stride into the fine-tuned DINO main encoder. Bold: best stride per encoder.

	DINOv2 (FT master)					DINOv3 (FT master)				
Class	base	OS4	OS8	OS16	OS32	base	OS4	OS8	OS16	OS32
person	26.1	31.1	32.4	35.8	27.9	30.4	30.3	26.6	34.4	29.2
rider	23.1	25.7	27.7	26.0	22.5	23.9	24.7	22.5	26.8	22.9
car	48.8	55.0	54.6	56.2	50.4	52.7	53.6	49.4	55.3	50.7
truck	38.2	38.3	37.9	36.8	35.8	32.5	38.6	36.9	40.6	36.0
bus	60.8	63.4	62.5	63.0	60.4	60.5	59.2	56.4	64.6	59.7
train	48.7	36.7	42.6	47.4	44.2	45.1	50.2	46.8	48.3	48.2
mbike	20.9	23.2	22.0	23.1	21.9	21.0	19.9	19.7	22.3	19.3
bicycle	17.2	20.2	20.0	21.6	17.5	18.5	19.3	15.6	20.1	16.7
mAP	35.5	36.7	37.5	**38.7**	35.1	35.6	37.0	34.2	**39.1**	35.3

Table 11. Computational cost at Cityscapes resolution (1024×2048) with a Mask2Former head, measured on a single A40 GPU.

Backbone	Params	GFLOPs	Throughput
DINOv2	108.1	1294	1.6
DINOv3	107.1	1083	2.2
SAM	111.1	2302	1.4
SAM2	89.1	1132	2.3
Ours-D3S2	176.2	1857	1.4

B.2. Efficiency and Computational Cost

Tab. 11 reports the computational cost of each backbone. Ours-D3S2 adds 774 GFLOPs over DINOv3-B (1857 vs. 1083) because we attach SAM2's backbone as an auxiliary edge provider. Despite the additional encoder, throughput remains comparable to single-encoder baselines such as DINOv2-B (1.4 vs. 1.6 img/s).

C. Additional Qualitative Results

Fig. 5 extends the main-paper comparison (Fig. 1) by including fine-tuned variants and our fused model. The two VFM families exhibit distinct failure patterns under

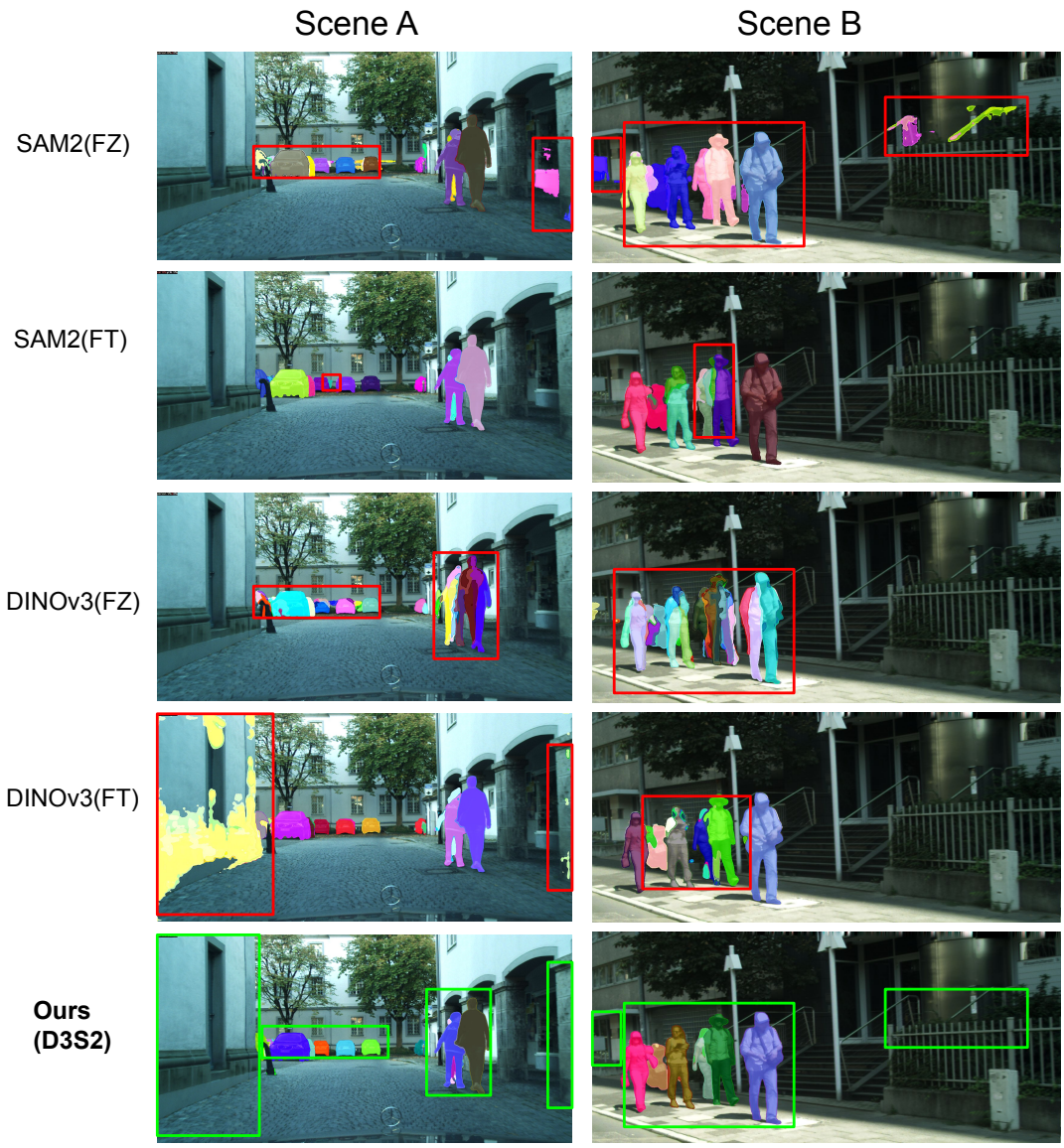

Figure 5. Qualitative comparison on Cityscapes. All outputs are unfiltered (no confidence thresholding). Red boxes: failure cases of individual encoders; green boxes: corresponding regions in ours-D3S2). Each column shows a different scene; rows show different encoder configurations.

both frozen and fine-tuned settings. Frozen SAM2 preserves tight boundaries but suffers from *category confusions*, missed small instances, and spurious fragments (e.g., distant regions in Scene A, staircase artifacts in Scene B). Frozen DINOv3 produces more coherent regions but tends to *over-segment* single objects into multiple instances (e.g., fragmented person in Scene A, split pedestrians in Scene B). Fine-tuning SAM2 slightly improves recall but introduces noisier masks; fine-tuning DINOv3 alleviates some fragmentation but causes new artifacts such as mask spillover into background regions (Scene A). Our fused model (D3S2) alleviates both failure modes by injecting edge-aware features from frozen SAM2 into the DINOv3 main encoder, yielding cleaner, better-separated, and more accurately localized predictions across both scenes.