TFM^2: Training-Free Mask Matching for Open-Vocabulary Semantic Segmentation Supplementary Materials

Yaoxin Zhuo[1]*, Zachary Bessinger[2], Lichen Wang[2], Naji Khosravan[2], Baoxin Li[1], Sing Bing Kang[2]
[1] Arizona State University, [2] Zillow Group
{yzhuo6, baoxin.li}@asu.edu
{zacharybe, lichenw, najik, singbingk}@zillowgroup.com

1. Details of the Predefined Sentence Templates

For mask classification, OVSS models use sentence templates to generate text embedding for each class in the target dataset. Following the common procedures in previous works [4, 8, 9], we fill the category names into these predefined sentence templates and then feed them into the text encoder of VLP models. We then average these text embeddings as the final text embedding $w_n \in \mathbb{R}^{1\times C}$ for category n. The templates are shown in Tab. 1.

While fixed sentence templates serve as a practical starting point, more advanced techniques in prompt learning [1–3,5–7,10–14] may help to further improve mask proposal classification performance. However, it is necessary to note that the focus of this work is not on these advanced techniques but rather on the fundamental process of text embedding generation using these predefined templates.

"a photo of a { }.",
"This is a photo of a { }",
"There is a { } in the scene",
"There is the { } in the scene",
"a photo of a { } in the scene",
"a photo of a small { }.",
"a photo of a medium { }.",
"a photo of a large { }.",
"This is a photo of a small { }.",
"This is a photo of a medium { }.",
"This is a photo of a large { }.",
"This is a small { } in the scene.",
"This is a medium { } in the scene.",
"This is a large { } in the scene.",

Table 1. Prompt templates of each category for text embeddings

2. Details of the mIoU comparison Figures and Tables.

Due to space constraints, we have condensed the presentation of performance figures for various OVSS methods using the TFM^2 on four datasets. Specifically, we have consolidated the figures for the 32-shot TFM^2 onlyin the table. However, for a more detailed analysis and a comprehensive view of the results, we kindly refer readers to Fig. 1 for the detailed figures and Tab. 2 for comprehensive tables. Notably, the performance of TFM^2 on the SAN [8] dataset still surpasses that on the SimSeg [9] and OVSeg [4] datasets in multiple shot settings. This discrepancy can be attributed to the ensemble weights of SimSeg and OVSeg.

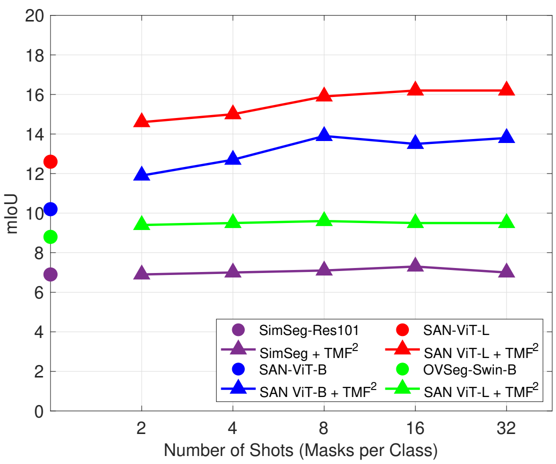
Additionally, it is essential to highlight that the performance of TFM^2 can be significantly affected by the number of reference masks available. This influence becomes particularly pronounced in cases with a limited number of reference masks per class, such as the 2-shot scenarios observed in the SimSeg on the PC-59 dataset, the OVSeg on the PC-59 dataset, and the SAN ViT-B on the PC-459 dataset. In these cases, the impact of reference mask numbers on TFM^2's performance may even lead to negative results. For a comprehensive understanding of these observations and their implications, we encourage readers to refer to the detailed figures and tables mentioned above.

3. Details of the Qualitative Results

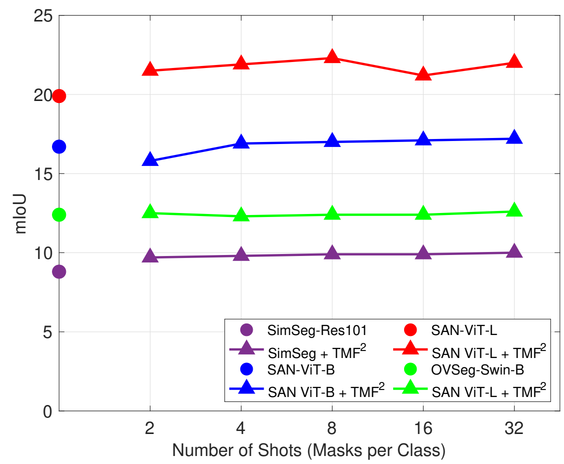
To improve visualization, larger images are included in Fig. 2. These images show detailed results, demonstrating TFM^2's role in aiding SAN's mask proposal classification. This figure further illustrates TFM^2's contribution to SAN's mask proposal classification.

However, it is worth noting that TFM^2 occasionally misclassified mask proposals as shown in Fig. 3. These figures highlight the importance of the quality of reference mask features (keys) used in TFM^2. Since the model heavily relies on the quality of these reference masks, low-quality or noisy references can lead to misclassification of TFM^2. For further research, it is valuable to explore methods for selecting high-quality reference masks when constructing the Mask Cache for TFM^2. This could contribute to improving the TFM^2's accuracy and robustness of mask proposal classification part in semantic segmentation task.

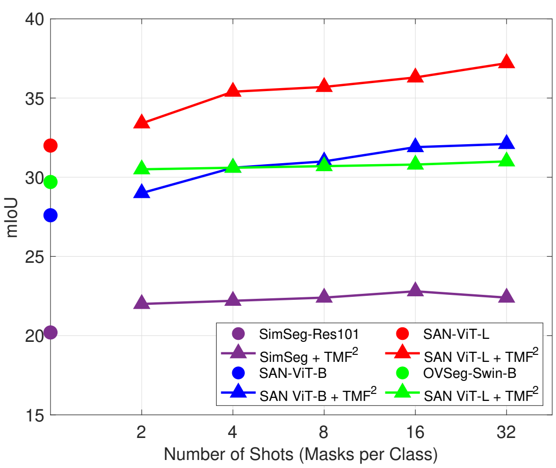
*Work was done while Yaoxin Zhuo was an intern at Zillow Group.



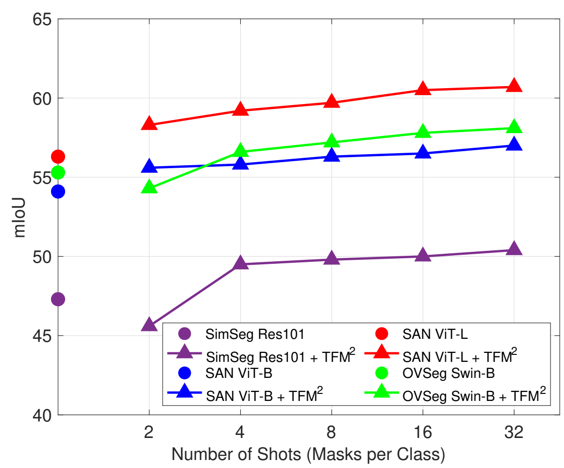
(a) ADE20K-847



(b) PC-459



(c) ADE20K-150



(d) PC-59

Figure 1. The mIoU of TFM^2 with varying number of shots and recent SOTA OVSS methods on four datasets.

Method	Pre-Trained Dataset	Ensemble	Shot Number	ADE-847	PC-459	ADE-150	PC-59
SimSeg (ECCV 2022)	COCO-Stuff	Yes	-	6.8	8.8	20.2	47.3
OVSeg (CVPR 2023)	COCO-Stuff	Yes	-	9.0	12.4	29.7	55.3
FC-CLIP (NeurIPS 2023)	COCO-Panoptic	Yes	-	14.8	18.2	34.1	58.4
ALIGN (ICML 2021)	-	No	-	4.8	5.8	12.9	22.4
GroupViT (CVPR 2022)	GCC + YFCC	No	-	4.3	4.9	10.6	25.9
Kunyang *et al.* (ICCV 2023)	COCO-Panoptic	No	-	3.5	7.1	18.8	45.2
OpenSeg (ECCV 2022)	COCO-Panoptic + COCO-Caption	No	-	6.8	11.2	24.8	45.9
MaskCLIP (ICML 2023)	COCO-Panoptic	No	-	8.2	10.0	23.7	45.9
SAN (CVPR 2023) (ViT-B)	COCO-Stuff	No	-	10.2	16.7	27.6	54.1
SAN (CVPR 2023) (ViT-L)	COCO-Stuff	No	-	12.6	19.9	32.0	56.3
ODISE (CVPR 2023)	COCO-Panoptic	No	-	11.1	14.5	29.9	57.3
DeOp (ICCV 2023)	COCO-Panoptic	No	-	7.1	9.4	22.9	48.8
MasQCLIP (ICCV 2023)	COCO-Panoptic	No	-	10.7	18.2	30.4	57.8
SimSeg (ResNet101)	COCO-Stuff	Yes	-	6.8	8.8	20.2	47.3
SimSeg (ResNet101) + **TFM^2**	COCO-Stuff	Yes	2	6.9(+0.1)	9.8(+1.0)	22.0(+1.8)	45.6(-1.7)
SimSeg (ResNet101) + **TFM^2**	COCO-Stuff	Yes	4	7.0(+0.2)	9.9(+1.1)	22.2(+2.0)	49.5(+2.2)
SimSeg (ResNet101) + **TFM^2**	COCO-Stuff	Yes	8	7.1(+0.3)	9.9(+1.1)	22.4(+2.2)	49.8(+2.5)
SimSeg (ResNet101) + **TFM^2**	COCO-Stuff	Yes	16	7.3(+0.4)	9.9(+1.1)	22.8(+2.6)	50.0(+2.7)
SimSeg (ResNet101) + **TFM^2**	COCO-Stuff	Yes	32	7.0(+0.2)	9.9(+1.1)	22.4(+2.2)	50.4(+3.1)
OVSeg (Swin-B)	COCO-Stuff	Yes	-	9.0	12.4	29.7	55.3
OVSeg (Swin-B) + **TFM^2**	COCO-Stuff	Yes	2	9.4(+0.4)	12.5(+0.1)	30.6(+0.9)	54.3(-1.0)
OVSeg (Swin-B) + **TFM^2**	COCO-Stuff	Yes	4	9.5(+0.5)	12.4(+0.0)	30.6(+0.9)	56.6(+1.3)
OVSeg (Swin-B) + **TFM^2**	COCO-Stuff	Yes	8	9.6(+0.6)	12.4 (+0.0)	30.7(+1.0)	57.2(+1.9)
OVSeg (Swin-B) + **TFM^2**	COCO-Stuff	Yes	16	9.5(+0.5)	12.4(+0.0)	30.8(+1.1)	57.8(+2.5)
OVSeg (Swin-B) + **TFM^2**	COCO-Stuff	Yes	32	9.5(+0.5)	12.6(+0.2)	31.0(+1.3)	58.1(+2.8)
SAN (ViT-B)	COCO-Stuff	No	-	10.2	16.7	27.6	54.1
SAN (ViT-B)+ **TFM^2**	COCO-Stuff	No	2	11.9(+1.7)	15.8(-0.9)	29.0(+1.4)	55.6 (+1.5)
SAN (ViT-B)+ **TFM^2**	COCO-Stuff	No	4	12.7(+2.5)	16.9(+0.2)	30.6(+3.0)	55.8 (+1.7)
SAN (ViT-B)+ **TFM^2**	COCO-Stuff	No	8	13.9(+3.7)	17.0(+0.3)	31.0(+3.4)	56.3 (+2.2)
SAN (ViT-B)+ **TFM^2**	COCO-Stuff	No	16	13.5(+3.3)	17.1(+0.4)	31.9(+4.3)	56.5 (+2.4)
SAN (ViT-B)+ **TFM^2**	COCO-Stuff	No	32	13.8(+3.6)	17.2(+0.5)	32.1(+4.5)	57.0 (+2.9)
SAN (ViT-L)	COCO-Stuff	No	-	12.6	19.9	32.0	56.3
SAN (ViT-L)+ **TFM^2**	COCO-Stuff	No	2	14.6(+2.0)	21.5(+1.6)	33.4(+1.4)	58.3 (+2.0)
SAN (ViT-L)+ **TFM^2**	COCO-Stuff	No	4	15.0(+2.4)	21.9(+2.0)	35.4(+3.4)	59.2 (+2.9)
SAN (ViT-L)+ **TFM^2**	COCO-Stuff	No	8	15.9(+3.3)	**22.3**(+2.4)	35.7(+3.7)	59.7 (+2.9)
SAN (ViT-L)+ **TFM^2**	COCO-Stuff	No	16	**16.2**(+3.4)	21.2(+1,3)	36.3(+4.3)	60.5 (+4.2)
SAN (ViT-L) + **TFM^2**	COCO-Stuff	No	32	**16.2**(+3.6)	22.0(+2.1)	**37.2**(+5.2)	**60.7**(+4.4)

Table 2. The mIoU comparison results of applying TFM^2 with different numbers of shots on multiple OVSS models.

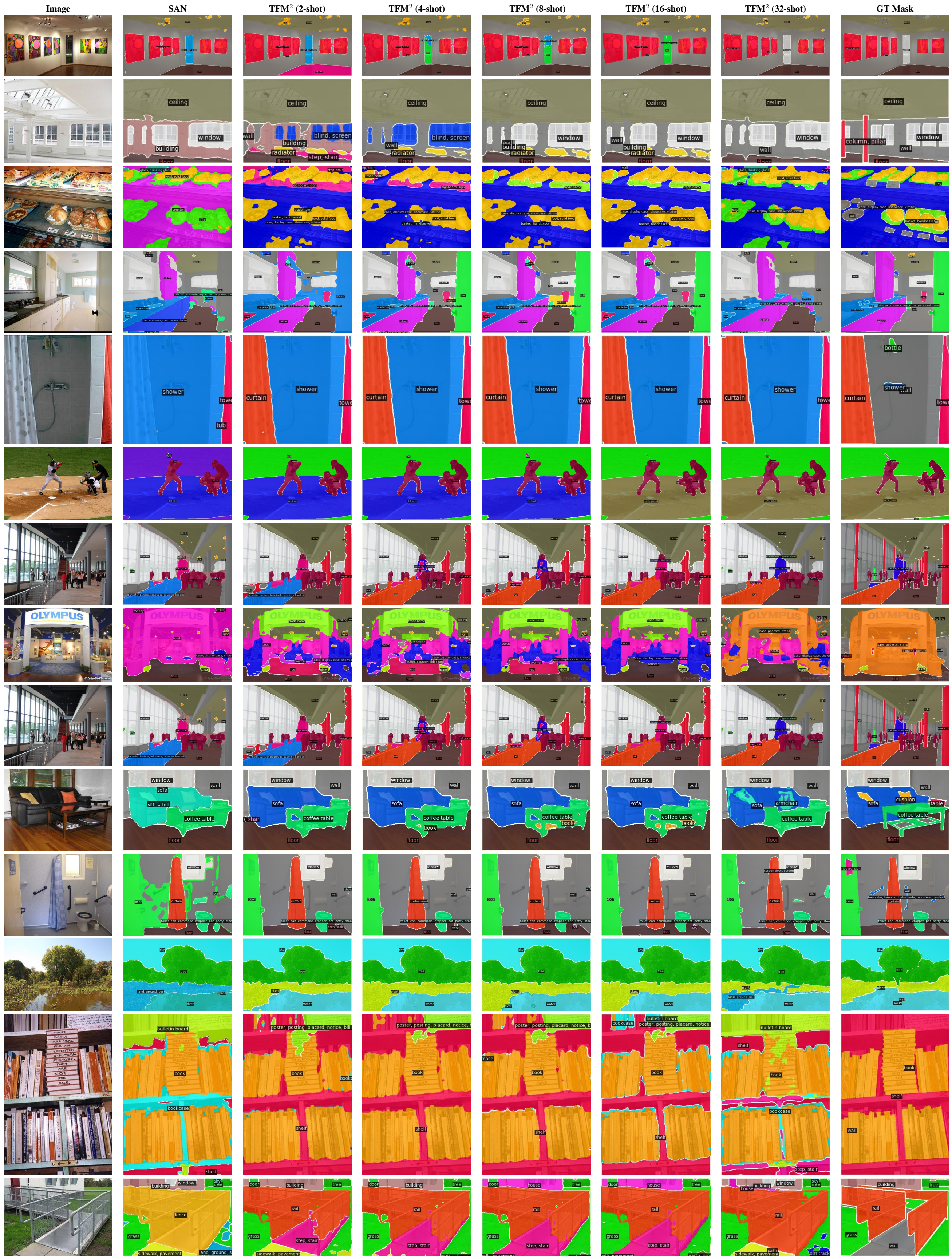


Figure 2. Qualitative examples showing TFM^2's role in improving mask proposal classification on ADE20k-150. The second column shows SAN inference without TFM^2. We see that SAN + TFM^2 (third column to seventh column) can steadily improve semantic segmentation when compared with the ground truth (last column). Please note that the color palette is the same for all mask classes.

Figure 3. Qualitative examples showing TFM^2 sometimes can not improve mask proposal classification on ADE20k-150. The second column shows SAN inference without TFM^2. We see that SAN + TFM^2 (third column to seventh column) sometimes misclassified some mask proposals (compared with the ground truth in the last column). Please note that the color palette is the same for all mask classes.

References

[1] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 1

[2] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 1

[3] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 1

[4] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 1

[5] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 35:14274–14289, 2022. 1

[6] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *NeurIPS*, 35:30569–30582, 2022. 1

[7] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, pages 5227–5237, 2022. 1

[8] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 1

[9] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, pages 736–753. Springer, 2022. 1

[10] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 1

[11] Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization. *arXiv preprint arXiv:2205.00049*, 2022. 1

[12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1

[13] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1

[14] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, pages 15659–15669, 2023. 1