Prompt-Calibrated SAM 3 for Open-Vocabulary Remote-Sensing Semantic Segmentation

Yanghui Song, Nanqing Liu, *Member, IEEE*, Haonan Yin, Yingjie Gao, Chengfu Yang, and Qi Ming

***Abstract*—Open-vocabulary semantic segmentation (OVSS) in remote-sensing images aims to segment categories beyond a fixed label space. Recent SAM3-based methods provide a promising training-free foundation; yet, three key issues remain: 1) a single class-name prompt lacks sufficient semantic coverage for complex remote-sensing categories; 2) expanding each category into multiple prompts introduces redundant online text encoding; and 3) directly aggregating multiple prompt responses propagates noisy activations into the final prediction. To address these issues, we propose ProC-SAM3, which calibrates SAM3's prompt interface for remote-sensing OVSS from three complementary aspects. First, we construct an offline prompt pool where a category matcher (CM) groups MLLM-generated candidates into per-category sets, and expansion constraints further refine each set using category-specific prior knowledge. Second, the resulting text embeddings are cached and reused across all test images, eliminating repeated text encoding. Third, we introduce presence-guided residual fusion (PGRF) to gate unreliable decoder outputs by prompt presence and confidence, followed by peak-preserving class aggregation that retains fine-grained activations for small and sparse objects. Experiments on eight benchmarks show that ProC-SAM3 achieves an average mIoU of 56.1%, outperforming the previous best training-free method by 3.9 percentage points. Code will be available at https://github.com/YanghuiSong/ProC-SAM3**

***Index Terms*—MLLMs, open-vocabulary semantic segmentation (OVSS), prompt calibration, remote-sensing, SAM3.**

I. Introduction

OPEN-VOCABULARY semantic segmentation (OVSS) in remote-sensing images aims to interpret complex geospatial scenes beyond a fixed set of predefined categories. Existing vision-language segmentation methods [1], [2], [3], [4] leverage CLIP-style patch-level features to transfer category semantics, but their lack of instance-aware segmentation capability limits performance on remote-sensing images [5], [6], [7]. The recent SAM3, a ViT-based foundation segmentation model, with its built-in promptable mask decoder, offers a fundamentally stronger segmentation backbone and has quickly surpassed CLIP-based methods on remote-sensing OVSS benchmarks [7], [8], [9], [10].

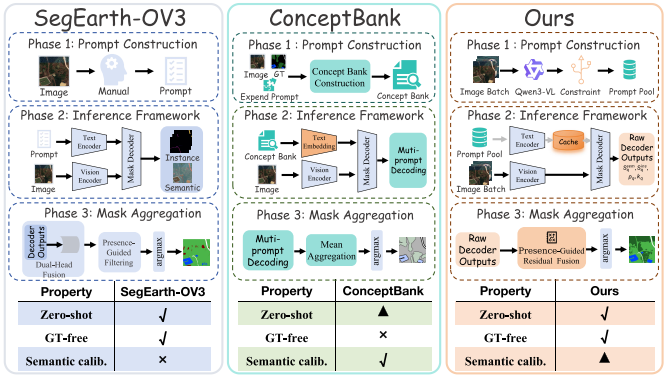

Fig. 1. Comparison of SAM3 adaptation paradigms for remote-sensing OVSS across three phases: prompt construction, inference, and mask aggregation. SegEarth-OV3 lacks semantic calibration and encodes prompts repeatedly online; conceptBank achieves calibration but relies on ground-truth annotations; our ProC-SAM3 provides GT-free calibration with cached embeddings and presence-guided aggregation.

However, the key challenge has shifted from segmentation capability to prompt calibration: how should SAM3's text-prompt interface be adapted to remote-sensing categories? Three issues arise in practice: 1) a single class-name prompt provides insufficient semantic coverage for visually diverse remote-sensing categories; 2) naively expanding each category into multiple prompts multiplies the online text-encoding cost; and 3) directly aggregating multiple prompt responses propagates noisy activations into the final prediction. As illustrated in Fig. 1, existing adaptation strategies address these issues to different extents across three phases: prompt construction, inference, and mask aggregation. SegEarth-OV3 [7] uses manually designed prompts without semantic calibration, encodes all prompts online for every image, and fuses decoder outputs via dual-head fusion with presence-guided filtering. While fully training-free and GT-free, it lacks domain-aware prompt calibration and incurs redundant text encoding. ConceptBank [9] constructs a semantically calibrated concept bank and caches text embeddings to avoid repeated encoding, but its concept bank is built from ground-truth masks, limiting its applicability to unannotated scenarios. Moreover, its mean aggregation of multiple prompt responses tends to suppress the activations of small objects.

To this end, we propose ProC-SAM3, a training-free and GT-free framework that improves SAM3 adaptation for remote-sensing OVSS across prompt construction, inference, and mask aggregation. Our contributions are as follows.

Received 2 May 2026; revised 19 June 2026; accepted 6 July 2026. Date of publication 14 July 2026; date of current version 23 July 2026. This work was supported in part by the Basic Research Project of Natural Science of Yunnan Province under Grant 202301AT070065 and Grant 202601AT070017. *(Corresponding author: Chengfu Yang.)*

Yanghui Song, Nanqing Liu, Haonan Yin, and Chengfu Yang are with the School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China (e-mail: yanghuisong55@gmail.com; lansing163@163.com; haonanyin26@gmail.com; yangchengfu@ynnu.edu.cn).

Yingjie Gao is with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: gaoyingjie@buaa.edu.cn).

Qi Ming is with the College of Computer Science, Beijing University of Technology, Beijing 100124, China (e-mail: chaser.ming@gmail.com).

Digital Object Identifier 10.1109/LGRS.2026.3713378

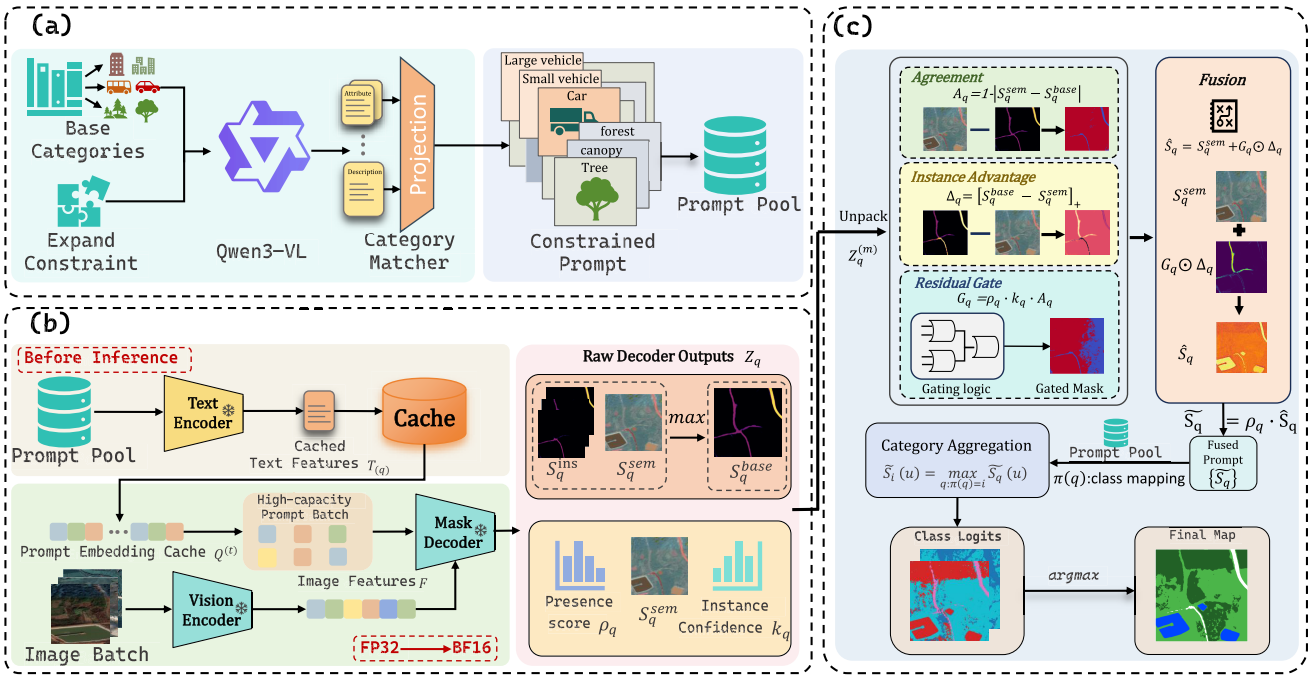

Fig. 2. Overview of the proposed ProC-SAM3, which operates in a fully training-free manner through three stages. (a) It first constructs a semantically enriched prompt pool offline via MLLM candidate generation, category matching, and expansion constraint refinement. (b) It then caches the prompt embeddings and performs efficient batched decoding. (c) It finally fuses per-prompt responses into per-class predictions through presence-guided residual gating and pixel-wise max-pooling aggregation.

1) We design offline prompt-pool construction (OPC) to calibrate category prompts without using ground-truth masks. OPC uses an MLLM to generate candidate descriptions and refines them through category matching and constraint-based projection, producing a compact prompt pool with clearer semantic boundaries.
2) We introduce prompt-optimized batched inference (POBI) to reduce the cost introduced by prompt expansion. POBI encodes the offline prompt pool once, caches the text embeddings, and reuses them during batched decoding, avoiding redundant online text encoding.
3) We propose presence-guided residual fusion (PGRF) for multiprompt mask aggregation. PGRF gates per-prompt decoder outputs by presence score and instance confidence and then applies pixelwise max-pooling to preserve fine-grained activations for small and sparse objects.
4) Experiments on eight remote-sensing OVSS benchmarks show that ProC-SAM3 achieves an average mIoU of 56.1%, outperforming the previous best training-free method by 3.9 percentage points.

II. Methodology

Given a remote-sensing image $x \in \mathbb{R}^{H\times W\times 3}$ and a category set $\mathcal{C} = \{c_1, \ldots, c_N\}$, our goal is to predict pixelwise semantic labels $\hat{Y} \in \{1, \ldots, N\}^{H\times W}$ without any task-specific training. As illustrated in Fig. 2, ProC-SAM3 improves SAM3's prompt interface from three complementary perspectives: *what to prompt*, where OPC enriches each category into a semantically diverse prompt set (Section II-A); *how to encode*, where POBI caches the prompt embeddings once and reuses them across all test images (Section II-B); and *how to aggregate*, where PGRF gates unreliable prompt responses and retains fine-grained activations through pixelwise max-pooling (Section II-C).

A. Offline Prompt-Pool Construction

A single class-name prompt often provides insufficient semantic coverage for complex remote-sensing categories because the same category may appear in different shapes, scales, and scene contexts. To address this, we build an offline prompt pool through a coarse-to-fine semantic refinement process, as illustrated in Fig. 2(a). In this process, each base category acts as a semantic anchor, and the final goal is to expand it into a compact set of descriptive prompts that are diverse enough to capture appearance variation but still stable enough to be reused at test time.

We first utilize Qwen3-VL to generate a diverse candidate set by jointly processing each base category c_i and its expansion constraints EC_i. The raw outputs are intentionally broad, containing both informative and potentially ambiguous candidates. To resolve ambiguity, we apply the Category Matcher (CM) function, which maps each candidate to its optimal base class. For example, "large vehicle," "small vehicle," and "car" are assigned to the vehicle class, while "forest," "canopy," and "tree" are grouped under trees. This step yields the refined category-specific candidate set $\widetilde{\mathcal{P}}_i$ for each c_i, formalized as

$$\widetilde{\mathcal{P}}_i = \mathrm{CM}\big(\mathrm{Qwen3}\,(c_i, EC_i)\big). \tag{1}$$

The matched candidates are then refined by category-specific expansion constraints, which encode prior knowledge about expected attributes and descriptions. In the figure, these constraints are summarized as attribute and description

cues. In practice, they can cover scale and material cues, object form, visual appearance, and scene environment, such as "large vehicle," "forest," "asphalt road," or "paved road." Candidates that violate these constraints are discarded, and only the descriptions that remain consistent with the intended category semantics are retained. This refinement step plays the role of a lightweight projection: it keeps the useful diversity produced by the MLLM while removing prompts that are semantically loose or visually implausible for the remote-sensing domain

$$\mathcal{P}_i = \text{Projection}\left(\widetilde{\mathcal{P}}_i,\ \text{Attribute}_i,\ \text{Description}_i\right). \tag{2}$$

All per-category sets are merged into a global prompt pool $\mathcal{P}^* = \bigcup_{i=1}^{N} \mathcal{P}_i$, where each prompt q retains its category label $\pi(q)$ (with $\pi(q) \in \{1, \ldots, N\}$ mapping the prompt to its associated category index). The resulting pool is therefore not just a list of category names, but a curated set of executable text prompts with clearer semantic boundaries. The entire construction is performed offline and adds no cost at test time. Constraint definitions and matching details are provided in our released code.[1]

B. Prompt-Optimized Batched Inference

Expanding each category from one prompt to many improves semantic coverage but also multiplies the text encoder calls per image in a conventional online inference pipeline. Since the global prompt pool $\mathcal{P}^*$ is constructed offline and remains fixed for all test images, repeatedly encoding the same textual instructions is unnecessary. We therefore reorganize the inference pipeline into a prompt-optimized batched form [Fig. 2(b)]. Before online segmentation, the entire prompt pool is mapped into the SAM3 text feature space once and stored as persistent prompt embeddings

$$T_{\{Q\}} = \text{TextEncoder}\left(\mathcal{P}^*\right) \tag{3}$$

where Q denotes the set of prompt indices. We use q to index an individual prompt and $\pi(q) \in \{1, \ldots, N\}$ to denote its associated category (consistent with the definition in Section II-A). This cache is independent of the image content and can be directly reused across datasets or image batches as long as the category set and prompt pool are unchanged.

During online inference, an image batch $\mathcal{I}$ is processed by the vision encoder once to obtain shared visual features $F = \text{VisionEncoder}(\mathcal{I})$. The cached prompt embeddings are split into chunks of size B. For the tth chunk, $T_{\mathcal{Q}^{(t)}} \subset T_{\{Q\}}$, the shared visual features and cached prompt features are jointly fed into the mask decoder

$$Z_q = \text{MaskDecoder}\left(F,\ T_q\right), \quad q \in \mathcal{Q}^{(t)}. \tag{4}$$

This formulation avoids repeated text encoding and enables batched decoding over multiple prompts and images.

For each prompt q, the SAM3 decoder produces four outputs after Sigmoid activation: semantic mask logits S_q^{sem}, instance mask logits S_q^{ins}, a presence score ρ_q indicating whether the prompted object exists in the image, and an instance confidence k_q. Among the instance masks, we select the one with the highest confidence k_q as the representative instance output S_q^{ins} and define a base response $S_q^{\text{base}}(u) = \max\left(S_q^{\text{sem}}(u), S_q^{\text{ins}}(u)\right)$ that retains the stronger activation at each pixel u. These outputs serve as inputs to the subsequent fusion stage.

C. Presence-Guided Residual Fusion

After decoding, each category has multiple prompt-level responses that must be combined into a single per-class prediction. Two problems arise in this step: mismatched prompts may produce confident masks on incorrect regions, and simply averaging responses across prompts tends to suppress local activations of small or narrow targets. To address both, we design PGRF as follows. The SAM3 decoder provides both a semantic branch and an instance branch for each prompt. When a prompt is genuinely present in the image and both branches agree, the instance branch can provide sharper boundary details that complement the semantic response. However, when a prompt is absent or the two branches disagree, the instance response is unreliable and should be suppressed. PGRF implements this logic through three quantities

$$A_q = 1 - \left|S_q^{\text{sem}} - S_q^{\text{base}}\right| \tag{5}$$

$$\Delta_q = \left[S_q^{\text{base}} - S_q^{\text{sem}}\right]_+ \tag{6}$$

$$G_q = \rho_q \cdot k_q \cdot A_q. \tag{7}$$

Here, A_q measures pixelwise agreement between the semantic and instance branches (values near 1 indicate high agreement), Δ_q captures the positive residual where the instance response exceeds the semantic response, and G_q is a gating factor that combines presence ρ_q, confidence k_q, and agreement A_q. The fused prompt-level response is

$$\widetilde{S}_q = \rho_q \cdot \left(S_q^{\text{sem}} + G_q \odot \Delta_q\right). \tag{8}$$

When the gate conditions are not met, G_q approaches zero, and the output reduces to $\rho_q \cdot S_q^{\text{sem}}$. After fusion, the refined responses are mapped back to their categories via $\pi(q)$ and aggregated by pixelwise max-pooling

$$\widetilde{S}_i(u) = \max_{q:\,\pi(q)=i} \widetilde{S}_q(u). \tag{9}$$

This retains the strongest local activation within each class, avoiding the suppression of small-object responses caused by averaging. Finally, an arg max across categories yields the prediction map $\hat{Y}(u) = \arg\max_i \widetilde{S}_i(u)$.

III. Experiments

A. Implementation Details

We evaluate on eight remote-sensing datasets: Vaihingen, Potsdam, LoveDA, OpenEarthMap, iSAID, UAVid, UDD5, and VDD. Input images are resized to 1008 × 1008 by the SAM3 processor. The offline prompt pool is constructed once per dataset using Qwen3-VL-8B under constraint files, and text embeddings are precomputed and cached by the prompt bank. During inference, prompts are processed in batches with a size of 1 for UDD5, VDD, and UAVid, and 4 for other datasets. The

[1] https://github.com/YanghuiSong/ProC-SAM3

TABLE I

QUANTITATIVE COMPARISON ON EIGHT REMOTE-SENSING OVSS BENCHMARKS. OEM DENOTES OPENEARTHMAP; * DENOTES REPRODUCED RESULTS, ALL OF WHICH ARE EVALUATED USING THE SEGEARTH-OV3 CLASS DEFINITIONS. ALL VALUES ARE MIOU (%), AND AVG. IS COMPUTED OVER THE REPORTED ENTRIES

Method	Backbone	Dataset								Avg.
		LoveDA	Potsdam	Vaihingen	iSAID	OEM	UDD5	VDD	UAVid	
VIP* *ICML'26* [11]	DINOv3 + dino.txt	31.2	23.0	30.5	23.5	20.7	35.2	52.5	19.9	29.6
SCLIP *ECCV'24* [12]	CLIP	30.4	39.6	35.9	16.1	29.3	38.7	37.9	31.4	32.4
ClearCLIP *ECCV'24* [1]	CLIP	32.4	42.0	36.2	18.2	31.0	41.8	39.3	36.2	34.6
ProxyCLIP *ECCV'24* [13]	CLIP+DINOv2	34.3	49.0	47.5	21.8	38.9	40.8	47.8	35.8	39.5
SegEarth-OV *CVPR'25* [5]	CLIP	36.9	48.5	29.1	21.7	40.3	50.6	45.3	42.5	39.2
RSCLIP* *J-STARS'26* [14]	CLIP	35.5	47.4	40.3	21.2	39.8	50.4	42.5	42.5	39.9
CorrCLIP *ICCV'25* [2]	CLIP+DINO+SAM2	36.9	51.9	47.0	25.5	32.9	46.1	47.3	38.3	40.7
ConInfer *CVPR Fingding'26* [15]	CLIP+DINOv3	39.3	50.0	31.4	20.1	**42.0**	46.9	50.3	46.4	40.8
ReAttnCLIP *CVPR'26* [16]	CLIP	37.0	48.7	29.9	23.2	41.1	53.7	49.7	44.0	40.9
ActiveSAM* *arXiv'26* [17]	SAM3	41.1	40.6	45.1	**28.3**	40.7	**72.1**	62.2	53.3	47.9
SegEarth-OV3* *arXiv'25* [7]	SAM3	**41.6**	**57.2**	**60.8**	26.0	41.0	71.6	**64.4**	**54.7**	**52.2**
Ours	SAM3	**46.1**	**58.8**	**64.5**	**29.5**	**48.9**	**73.4**	**67.5**	**60.0**	**56.1**

image encoder operates in half precision, while the decoder remains in full precision. All experiments are conducted without training on a single RTX 4090 GPU (24 GB), with Qwen3 unloaded from GPU memory before SAM3 inference to avoid memory contention. Unless otherwise specified, quantitative results are reported as mIoU (%).

B. Comparison With State-of-the-Art Methods

Table I compares representative open-vocabulary segmentation methods on eight remote-sensing benchmarks. Since several prior methods do not report results on all datasets and the average is computed over available entries, the fairest direct comparison is with training-free methods that provide full coverage, especially SegEarth-OV3. Under this setting, our method achieves the highest mIoU on all eight benchmarks and improves the average mIoU from 52.2 to 56.1 (+3.9). The gains are consistent across datasets and are most notable on OpenEarthMap (+7.9%), UAVid (+5.3%), LoveDA (+4.5%), and Vaihingen (+3.7%). These datasets contain diverse land-cover patterns, dense small objects, and large appearance variations, suggesting that calibrated prompts are more effective than directly using category names or unconstrained prompt expansion.

The qualitative results in Fig. 3 show a similar trend. Compared with SegEarth-OV3, our method recovers more complete structures, generates cleaner land-cover regions, and retains stronger small-object responses. These improvements indicate that offline image-conditioned prompt calibration and presence-guided mask aggregation reduce semantic ambiguity while preserving fine-grained activations in complex remote-sensing scenes.

C. Ablation Study

1) Component Ablation: As shown in Table II, OPC provides the dominant gain (+3.7% avg. mIoU), indicating that the baseline bottleneck is semantic alignment rather than decoder capacity. PGRF further improves the average to 56.1, with consistent gains across all datasets. Fig. 4 extends this analysis by comparing PGRF against two alternative fusion strategies (max and mean) across three datasets. Max and mean achieve comparable performance (55.1 versus 54.4),

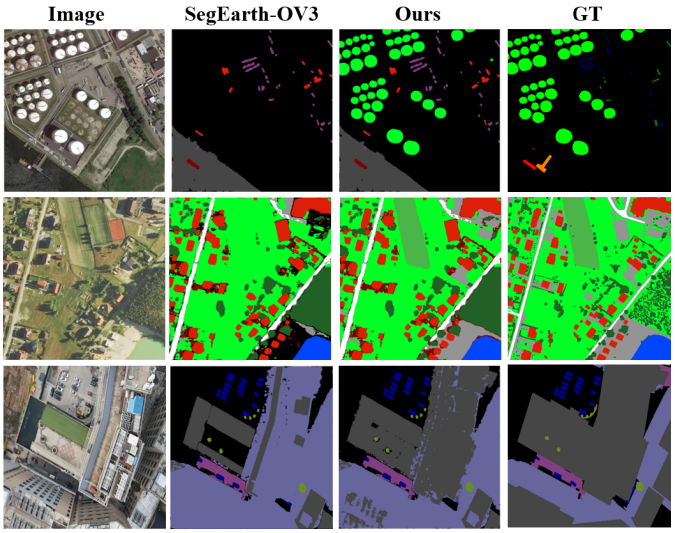

Fig. 3. Qualitative comparison on representative remote-sensing scenes.

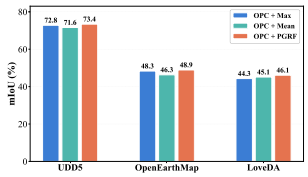

Fig. 4. Ablation on different fusion strategies.

TABLE II

ABLATION STUDY OF FRAMEWORK COMPONENTS ON FOUR REMOTE-SENSING DATASETS. PC DENOTES PROMPT CALIBRATION AND PGRF DENOTES PROMPT-GUIDED RESIDUAL FUSION

OPC	PGRF	UDD5	OEM	LoveDA	Avg.
		71.6	41.0	41.6	51.4
✓		72.8	48.3	44.3	55.1
✓	✓	**73.4**	**48.9**	**46.1**	**56.1**

while PGRF outperforms both, confirming its effectiveness in suppressing unreliable instance activations. Fig. 5 shows that POBI consistently accelerates inference across all datasets, with the largest gain observed on Potsdam (1.35 → 2.42 FPS).

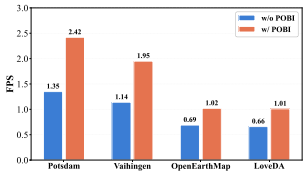

Fig. 5. Inference speed comparison with and without POBI.

TABLE III
ABLATION ON PROMPT CONSTRUCTION STRATEGIES ACROSS THREE DATASETS. ALL VARIANTS USE THE SAME INFERENCE PIPELINE

Prompt Strategy	UDD5	OEM	LoveDA	Avg.
Class-name prompt	71.6	41.0	41.6	51.4
Open-ended MLLM prompt	66.4	31.3	34.6	44.1
Text-only expansion	72.3	40.6	42.8	51.9
Image-conditioned expansion	72.7	45.1	44.3	54.0
Constraint-guided prompt pool	**73.4**	**48.9**	**46.1**	**56.1**

2) Prompt Construction Strategies: We further evaluate how different prompt construction strategies affect semantic alignment. Table III compares five strategies under the same inference pipeline: class-name prompts, open-ended MLLM prompts, text-only expansion, image-conditioned expansion, and the proposed constraint-guided prompt pool. To make the difference more concrete, we take the *road* class in UDD5 as an example. The class-name baseline uses only "*road*." The open-ended MLLM setting often produces broad variants such as "*road*, *street*, *path*, *highway*, *avenue*," while text-only expansion narrows them to words such as "*road*, *path*, *paved*." After introducing image context, the image-conditioned expansion yields more scene-aware phrases such as "*asphalt road*, *paved road*, *school road*." Our constraint-guided prompt pool further filters these candidates and retains a stable subset, e.g., "*road*, *asphalt road*, *paved road*," thereby preserving prompt diversity without causing semantic drift.

The quantitative results in Table III are consistent with the above examples. Open-ended MLLM prompts degrade performance on all three datasets, dropping the average mIoU from 51.4 to 44.1, which indicates that unconstrained expansion introduces substantial semantic drift. Text-only expansion is more stable and slightly improves the average to 51.9, but its gains remain limited because lexical variation alone cannot reliably capture the remote-sensing scene context. After introducing image context, the average mIoU further rises to 54.0, with a particularly clear gain on OEM (40.6 $\rightarrow$ 45.1), showing that scene-aware prompt generation is already beneficial. The proposed constraint-guided prompt pool achieves the best results on all three datasets, reaching 73.4 on UDD5, 48.9 on OEM, and 46.1 on LoveDA. Compared with the class-name baseline, this corresponds to gains of 1.8, 7.9, and 4.5 points, respectively, indicating that the combination of image conditioning and category-specific constraints yields the most reliable prompt set.

3) Efficiency of POBI: Fig. 5 compares the inference throughput without and with POBI. The figure shows consistent gains on all four datasets, from 1.14 to 1.95 on Vaihingen, 0.69 to 1.02 on OpenEarthMap, 0.66 to 1.01 on LoveDA, and 1.35 to 2.42 on Potsdam. This gain mainly comes from avoiding repeated text encoding and replacing fp32 inference with bf16, which reduces both computation and memory overhead during decoding.

IV. CONCLUSION

We presented ProC-SAM3, a training-free framework that improves SAM3's prompt interface for remote-sensing OVSS from three aspects: prompt semantics (OPC), encoding efficiency (POBI), and mask aggregation (PGRF). Experiments on eight benchmarks show consistent improvements over existing methods, with an average mIoU of 56.1%. Ablation results indicate that OPC contributes the largest accuracy gain, while PGRF and POBI provide complementary improvements in prediction quality and inference efficiency, respectively.

REFERENCES

[1] M. Lan, C. Chen, Y. Ke, X. Wang, L. Feng, and W. Zhang, "Clearclip: Decomposing clip representations for dense vision-language inference," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2024, pp. 143–160.

[2] D. Zhang, F. Liu, and Q. Tang, "CorrCLIP: Reconstructing patch correlations in CLIP for open-vocabulary semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 24677–24687.

[3] Q. Cao, Y. Chen, C. Ma, and X. Yang, "Open-vocabulary high-resolution remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, 2025.

[4] Z. Hu, Q. Xu, Y. Duan, Y. Tai, and H. Li, "Sota: Self-adaptive optimal transport for zero-shot classification with multiple foundation models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jan. 2026, pp. 26624–26634.

[5] K. Li et al., "SegEarth-OV: Towards training-free open-vocabulary segmentation for remote sensing images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 10545–10556.

[6] K. Li et al., "Annotation-free open-vocabulary segmentation for remote-sensing images," 2025, *arXiv:2508.18067*.

[7] K. Li et al., "SegEarth-OV3: Exploring SAM 3 for open-vocabulary semantic segmentation in remote sensing images," 2025, *arXiv:2512.08730*.

[8] N. Liu, X. Xu, Y. Su, H. Zhang, and H.-C. Li, "PointSAM: Pointly-supervised segment anything model for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025.

[9] G. Pei, X. Jiang, Y. Yao, X. Shu, F. Shen, and B. Jeon, "Taming SAM3 in the wild: A concept bank for open-vocabulary segmentation," 2026, *arXiv:2602.06333*.

[10] N. Carion et al., "SAM 3: Segment anything with concepts," 2025, *arXiv:2511.16719*.

[11] H. Zhu et al., "VIP: Visual-guided prompt evolution for efficient dense vision-language inference," 2026, *arXiv:2605.12325*.

[12] F. Wang, J. Mei, and A. Yuille, "Sclip: Rethinking self-attention for dense vision-language inference," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2024, pp. 315–332.

[13] M. Lan, C. Chen, Y. Ke, X. Wang, L. Feng, and W. Zhang, "Proxyclip: Proxy attention improves clip for open-vocabulary segmentation," in *Proc. Eur. Conf. Comput. Vis. Cham*, Switzerland: Springer, 2024, pp. 70–88.

[14] S. Wang, X. Sun, J. Han, and X. X. Zhu, "RSCLIP for training-free open-vocabulary remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 19, pp. 15626–15639, 2026.

[15] W. Chen, Z. Hu, Y. Zhang, H. Ning, and Y. Tai, "Coninfer: Context-aware inference for training-free open-vocabulary remote sensing segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Feb. 2026, pp. 7408–7418.

[16] X. Niu, M. Zhao, D. Jiang, Y. Wu, and B. Su, "ReAttnCLIP: Training-free open-vocabulary remote sensing image segmentation via re-defined attention in clip," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2026, pp. 24980–24989.

[17] T. D. Tien and Z. Shen, "ActiveSAM: Image-conditional class pruning for fast and accurate open-vocabulary segmentation," 2026, *arXiv:2606.16996*.