南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Test-Time Adaptation of Vision-Language Models for Open-Vocabulary Semantic Segmentation

Mehrdad Noori* David Osowiechi* Gustavo A. Vargas Hakim Ali Bahri
Moslem Yazdanpanah Sahar Dastani Farzad Beizaee Ismail Ben Ayed
Christian Desrosiers

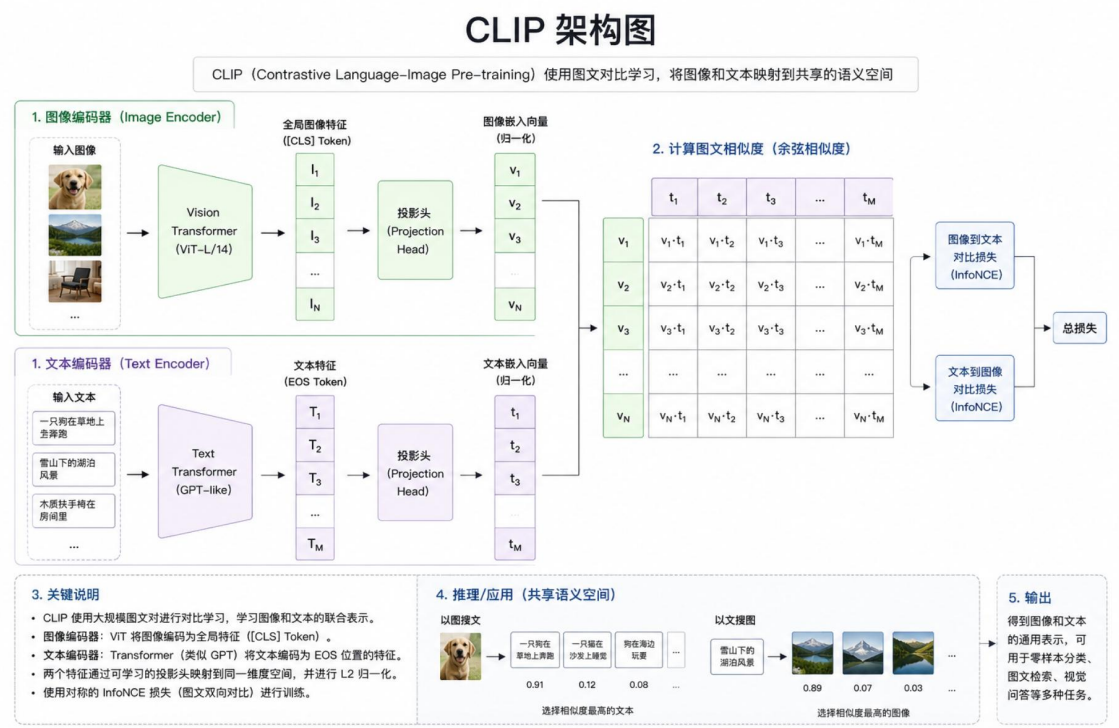
汇报人：蒋明忠

时间：2026.06

NeurIPS

Background

现有的OVSS方法已经取得了重大进展，但它们在测试时仍然容易受到域转移的影响，例如环境变化或图像损坏，这可能会大大降低分割质量。在缺乏使它们能够适应未见过的测试时间分布的机制的情况下，这些模型可能会失去泛化能力，从而限制了它们在实际应用中的可靠性。

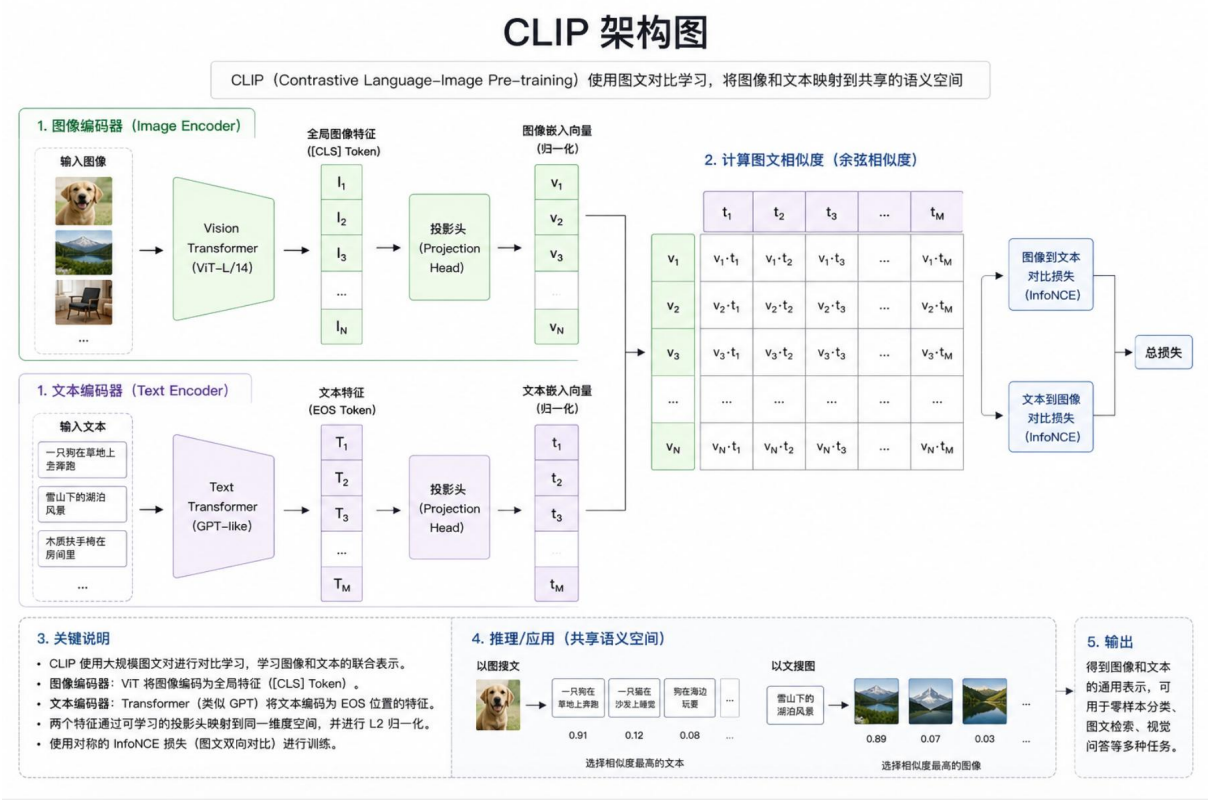


$$\arg\max_k \; sim(\mathbf{f}_{[\mathtt{cls}]}, \mathbf{t}_k), \quad \text{where } sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

Background

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

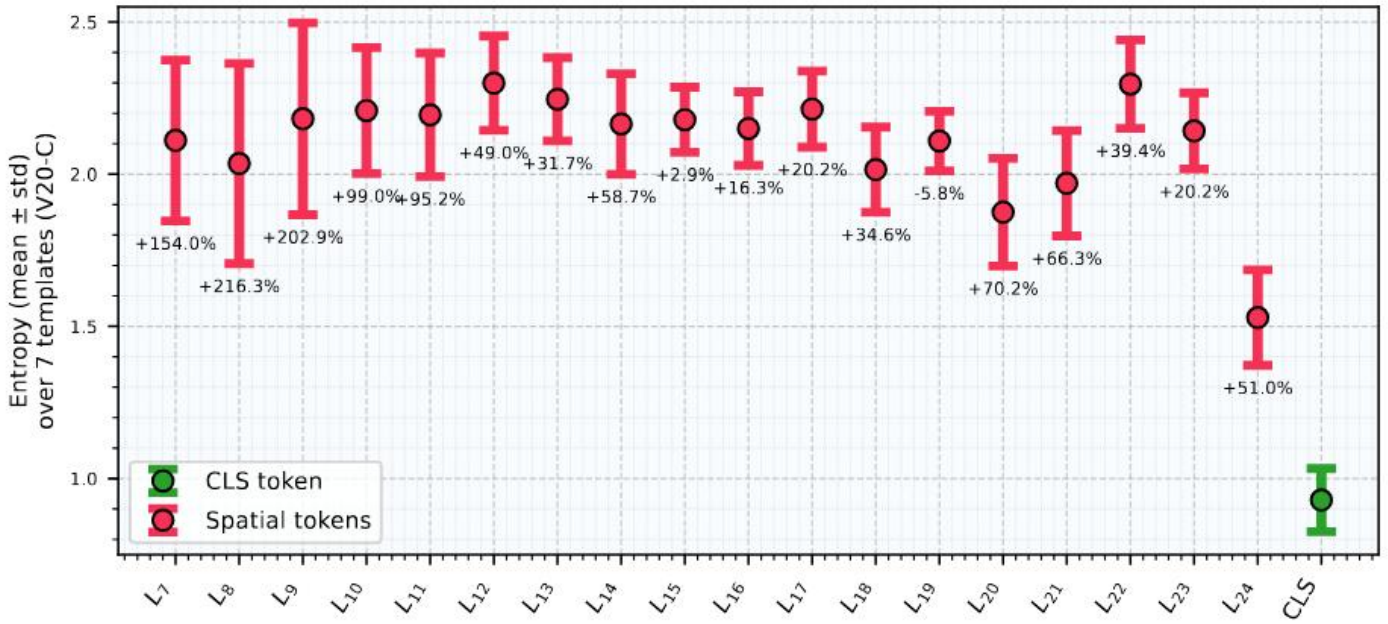
最近，测试时间自适应在图像分类的视觉语言模型中引起了广泛的兴趣。然而，据我们所知，在密集预测任务中，如开放词汇语义分割(OVSS)，这个问题完全被忽视了。为此，我们提出了一种新的TTA方法，以适应VLMs在测试期间进行分割。

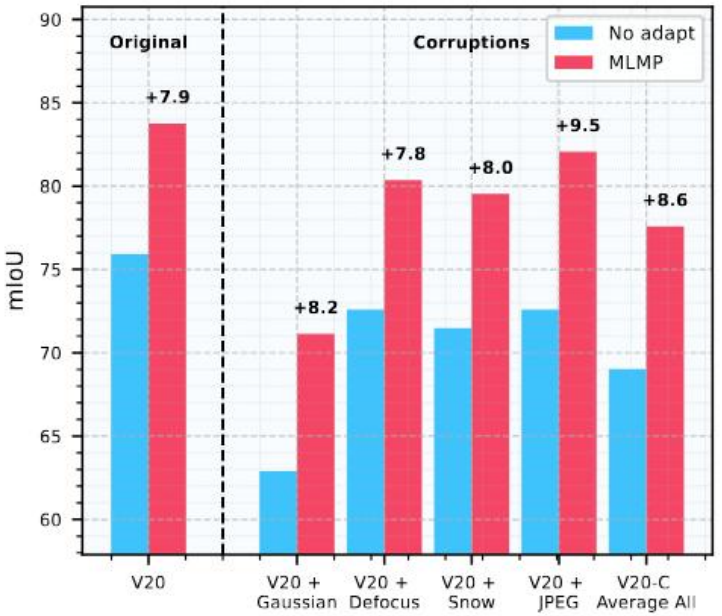


Method

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

与用于图像分类的TTA方法不同，我们的多层次和多提示(MLMP)熵最小化集成了来自中间视觉编码器层的特征，并在全局CLS标记和局部像素级使用不同的文本提示模板执行。我们的方法可以作为任何分割网络的即插即用，不需要额外的训练数据或标签，即使使用单个测试样本也保持有效。

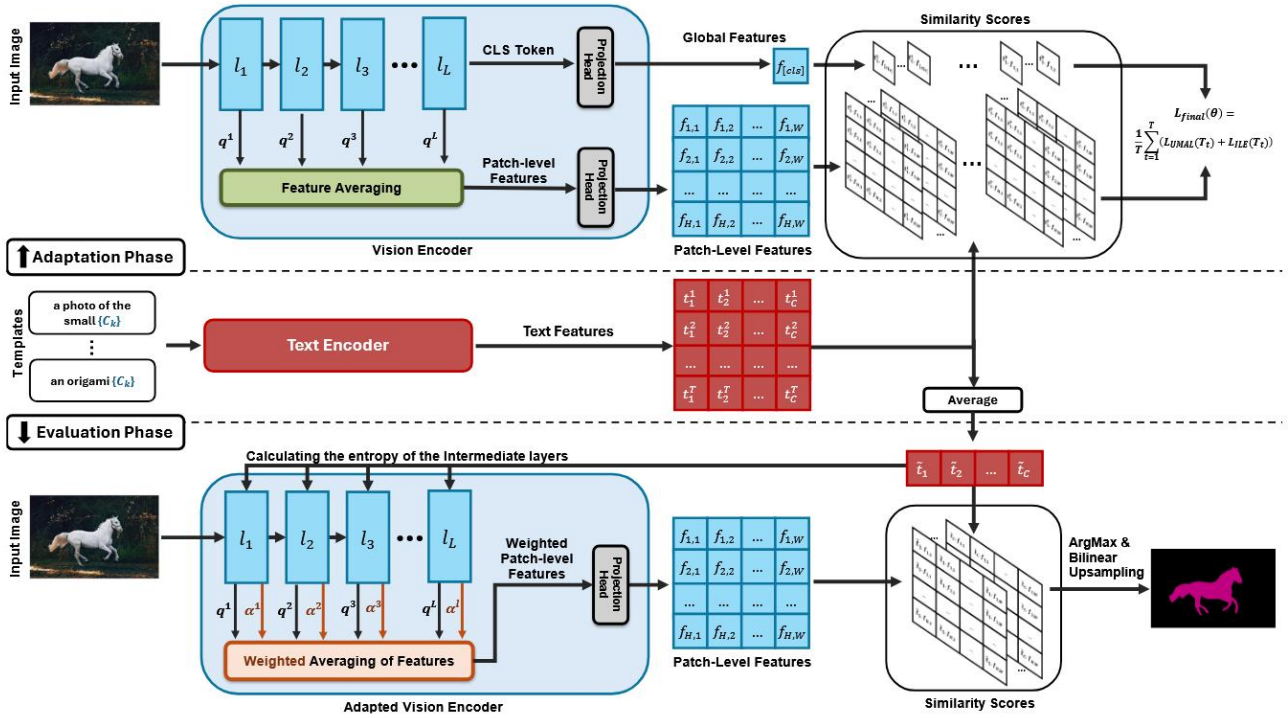




Method

自适应多层次融合：MLMP集成了中间视觉编码器层的功能，以捕获互补的、位移弹性的线索。为了进一步增强鲁棒性，我们提出了一种不确定性感知策略，该策略根据预测熵 重新加权来自各个层的特征。

多提示自适应：MLMP通过在全局CLS令牌和本地像素级直接最小化跨不同文本提示模板的熵,将提示灵敏度转换为信号。这种优化通过强制跨语言视角和特征深度的一致性,自然地补充了我们的多层次特征融合。



Method

自适应多层次融合：MLMP集成了中间视觉编码器层的功能，以捕获互补的、位移弹性的线索。为了进一步增强鲁棒性，我们提出了一种不确定性感知策略，该策略根据预测熵 重新加权来自各个层的特征。

不固定使用最后一层，而是根据当前测试图像自动判断哪些层更可靠，并把它们融合起来。

浅层：边缘、纹理、局部细节更强。中层：结构信息更明显。深层：语义信息更强。

不同腐蚀或 domain shift 下，不同层受影响程度不同。

$$\arg\max_k \ sim\left(\mathbf{f}_{[\mathtt{cls}]}, \mathbf{t}_k\right), \ \text{where } sim(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}.$$

$$p_{ik} = \frac{\exp(\mathrm{sim}(f_i, t_k)/\tau)}{\sum_{k'=1}^{|C|}\exp(\mathrm{sim}(f_i, t_{k'})/\tau)}$$

$$\mathbf{P} = \mathrm{softmax}\left(\mathrm{norm}(\mathbf{F})\cdot\mathrm{norm}(\mathbf{T})^\top/\tau\right).$$

$$\mathcal{H}(P) = -\frac{1}{B\cdot N}\sum_{i=1}^{B\cdot N}\sum_{k=1}^{|C|} p_{ik}\log p_{ik}$$

$$h^\ell = \mathcal{H}(\mathbf{P}^\ell), \ \text{with } \mathbf{P}^\ell = \mathrm{softmax}\left(\mathrm{norm}(\mathbf{F}^\ell)\cdot\mathrm{norm}(\mathbf{T})^\top/\tau\right).$$

$$\alpha^\ell = \frac{\exp(-\beta\cdot h^\ell)}{\sum_{\ell'=1}^{L}\exp(-\beta\cdot h^{\ell'})}.$$

$$\overline{\mathbf{F}} = \mathrm{proj}(\overline{\mathbf{Q}}), \ \text{with } \overline{\mathbf{Q}} = \sum_{\ell=1}^{L}\alpha^\ell\mathbf{Q}^\ell,$$

$$\mathcal{L}_{\mathrm{UAML}}(\mathbf{T}) = \mathcal{H}(\overline{\mathbf{P}}), \ \text{with } \overline{\mathbf{P}} = \mathrm{softmax}\left(\mathrm{norm}(\overline{\mathbf{F}})\cdot\mathrm{norm}(\mathbf{T})^\top/\tau\right).$$

$$\mathcal{L}_{\mathrm{ILE}}(\mathbf{T}) = -\frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{|C|} p_{b,k}^{[\mathtt{cls}]}\log p_{b,k}^{[\mathtt{cls}]}.$$

Method

多提示自适应：MLMP通过在全局CLS令牌和本地像素级直接最小化跨不同文本提示模板的熵,将提示灵敏度转换为信号。这种优化通过强制跨语言视角和特征深度的一致性,自然地补充了我们的多层次特征融合。

之前关于分类的TTA工作已经显示了利用VLM中的多个提示模板来编码类标签的有用性，基于模板捕获关于这些类的互补信息的想法。使用多个模板起到了跨模态正则化的作用，鼓励了更稳定和广义的学习信号。

对每个 prompt 模板都计算一次 patch-level 和 image-level 熵损失，然后对所有 prompt 的损失取平均，用这个平均损失更新视觉编码器参数。

论文认为，平均文本 embedding 会抹掉不同 prompt 的差异。而在 loss 层面分别计算每个 prompt 的损失，可以让每个 prompt 都成为一个独立的 “critic”。

$$\mathcal{L}_{\text{UAML}}(\mathbf{T}) = \mathcal{H}(\overline{\mathbf{P}}), \quad \text{with } \overline{\mathbf{P}} = \text{softmax}\big(\text{norm}(\overline{\mathbf{F}}) \cdot \text{norm}(\mathbf{T})^{\top}/\tau\big).$$

$$\mathcal{L}_{\text{ILE}}(\mathbf{T}) = -\frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{|\mathcal{C}|} p_{b,k}^{[\texttt{cls}]} \log p_{b,k}^{[\texttt{cls}]}.$$

$$\mathcal{L}_{\text{final}}(\theta) = \frac{1}{T}\sum_{t=1}^{T}(\mathcal{L}_{\text{UAML}}(\mathbf{T}_t) + \mathcal{L}_{\text{ILE}}(\mathbf{T}_t)).$$

Dataset: V20	Text	Params	Loss
Original	78.91 ±0.07	74.46 ±0.21	**79.70** ±0.06
Gaussian Noise	66.27 ±0.00	62.83 ±0.04	**66.75** ±0.01
Defocus Blur	74.05 ±0.10	70.28 ±0.16	**74.31** ±0.09
Snow	73.78 ±0.02	70.10 ±0.30	**74.66** ±0.01
JPEG Compression	74.98 ±0.05	70.55 ±0.11	**75.56** ±0.02
V20-C Average	71.92	68.44	**72.58**

Experiments

Table 2: mIoU performance when using different layer ranges in the proposed multi-level adaptation.

ViT-L/14 Layer Range	L_{24} (last)	L_{23-24} (last two)	L_{22-24} (last three)	L_{19-24} (last 25%)	L_{13-24} (last 50%)	L_{7-24} (last 75%)	L_{1-24} (all layers)
V20 (Original)	77.00 ±0.04	77.65 ±0.02	77.66 ±0.09	80.61 ±0.05	80.50 ±0.03	**81.67 ±0.04**	78.79 ±0.02
Gaussian Noise	63.02 ±0.06	64.41 ±0.05	65.39 ±0.13	66.88 ±0.18	66.88 ±0.02	**67.82 ±0.01**	63.06 ±0.09
Defocus Blur	72.06 ±0.12	72.93 ±0.19	72.84 ±0.02	76.10 ±0.16	76.37 ±0.05	**78.78 ±0.02**	77.56 ±0.09
Snow	71.04 ±0.05	72.09 ±0.04	72.56 ±0.02	74.47 ±0.12	74.41 ±0.01	**76.39 ±0.02**	73.72 ±0.07
JPEG Compression	71.84 ±0.15	73.88 ±0.11	74.40 ±0.07	76.96 ±0.02	77.67 ±0.03	**78.73 ±0.08**	75.87 ±0.19
V20-C Average	69.33	70.33	70.78	72.89	73.45	**74.90**	72.02

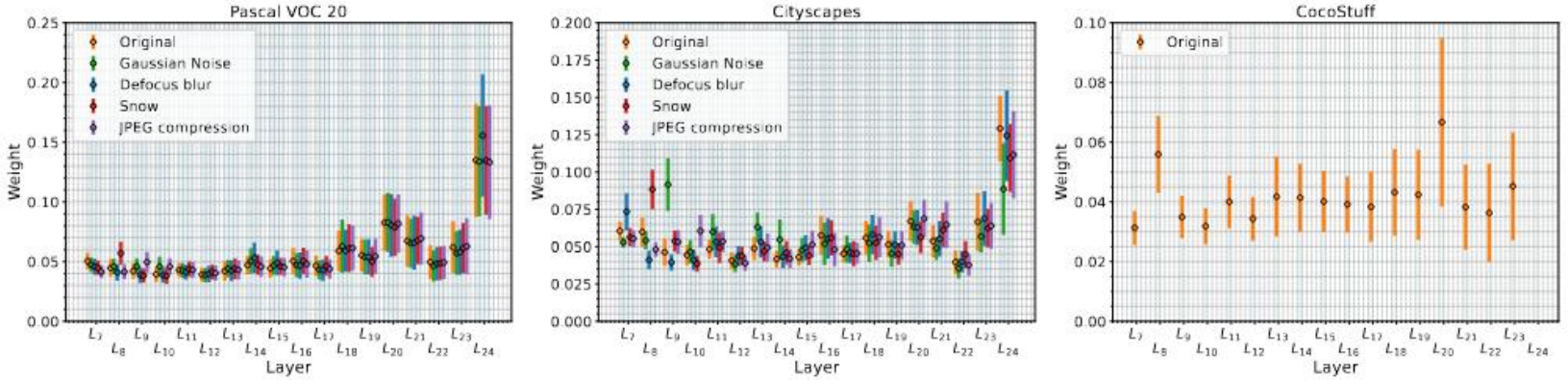


Figure 3: Mean and standard deviation of layer-wise confidence weights of MLMP across datasets. The fusion mechanism adaptively emphasizes more reliable layers based on input conditions.

Experiments

OVSS Backbone: NACLIP		Adaptation Method					
Dataset		No Adapt.	TENT	TPT	WATT	CLIPArTT	MLMP
V20 (Original)		75.91	77.00 ±0.04	75.93 ±0.01	57.73 ±0.06	72.77 ±0.14	**83.76 ±0.00**
V20-C	Gaussian Noise	62.89	63.02 ±0.06	62.98 ±0.01	36.44 ±0.04	53.36 ±0.25	**71.13 ±0.09**
	Shot noise	66.26	65.88 ±0.06	66.33 ±0.02	40.95 ±0.05	58.15 ±0.28	**75.02 ±0.03**
	Impulse Noise	63.16	64.17 ±0.04	63.12 ±0.01	34.90 ±0.06	54.83 ±0.03	**71.34 ±0.11**
	Defocus blur	72.59	72.06 ±0.12	72.55 ±0.02	52.43 ±0.03	65.39 ±0.45	**80.36 ±0.06**
	Glass blur	71.44	70.74 ±0.07	71.40 ±0.01	49.96 ±0.05	64.62 ±0.13	**78.84 ±0.05**
	Motion blur	73.10	73.50 ±0.10	73.16 ±0.02	53.35 ±0.06	67.48 ±0.17	**81.41 ±0.05**
	Zoom blur	59.03	61.36 ±0.07	59.00 ±0.01	41.39 ±0.08	52.37 ±0.12	**69.41 ±0.12**
	Snow	71.49	71.04 ±0.05	71.44 ±0.01	51.18 ±0.06	66.97 ±0.02	**79.53 ±0.05**
	Frost	65.38	67.01 ±0.02	65.46 ±0.01	45.75 ±0.05	60.48 ±0.08	**73.20 ±0.07**
	Fog	70.69	70.54 ±0.07	70.70 ±0.01	52.96 ±0.04	67.85 ±0.10	**79.81 ±0.06**
	Brightness	74.95	75.61 ±0.02	74.95 ±0.01	55.82 ±0.05	71.52 ±0.14	**83.51 ±0.01**
	Contrast	71.51	70.51 ±0.04	71.49 ±0.02	50.74 ±0.06	66.01 ±0.06	**79.06 ±0.16**
	Elastic transform	62.86	65.78 ±0.05	62.95 ±0.01	45.45 ±0.04	60.41 ±0.10	**74.03 ±0.01**
	Pixelate	77.28	76.95 ±0.12	77.31 ±0.01	59.76 ±0.05	73.14 ±0.17	**84.97 ±0.04**
	JPEG compression	72.59	71.84 ±0.15	72.56 ±0.01	53.44 ±0.05	68.21 ±0.07	**82.06 ±0.01**
	Average	69.01	69.33	69.03	48.30	63.39	**77.58**
V21 (Original)		45.12	45.65 ±0.02	45.17 ±0.01	28.58 ±0.05	39.50 ±0.04	**50.78 ±0.02**
V21-C Average		40.75	40.95	40.77	24.12	34.16	**46.25**
P59 (Original)		28.23	28.73 ±0.02	28.26 ±0.01	16.55 ±0.04	24.60 ±0.03	**31.95 ±0.02**
P59-C Average		23.88	23.88	23.88	13.37	19.72	**27.03**
P60 (Original)		24.95	25.29 ±0.01	24.98 ±0.01	14.77 ±0.03	21.88 ±0.03	**27.99 ±0.03**
P60-C Average		21.39	21.25	21.49	12.08	17.79	**24.07**
CityScapes (Original)		29.49	30.54 ±0.04	29.57 ±0.01	20.77 ±0.06	–	**33.35 ±0.03**
CityScapes-C Average		21.63	21.64	21.60	13.45	–	**23.02**
COCOObject (Original)		23.80	24.88 ±0.01	23.84 ±0.01	14.14 ±0.06	21.34 ±0.03	**28.84 ±0.01**
COCOStuff (Original)		18.34	18.76 ±0.01	18.35 ±0.01	9.49 ±0.02	15.48 ±0.01	**21.25 ±0.01**

Experiments

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 3: mIoU comparison of MLMP components, showing individual and combined contributions.

Multi-Level Fusion	✗	✓	✓	✗	✗	✓	✓	✓	✓	✗	✓	✓
Multi-Prompt Loss	✗	✗	✗	✓	✗	✓	✓	✗	✗	✓	✓	✓
Image-Level Entropy	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓
Uncertainty-Aware Weighting	✗	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓
V20 (Original)	77.00	77.38	81.67	79.70	78.74	78.97	83.00	77.69	82.70	81.15	79.13	**83.76**
Gaussian Noise	63.02	65.42	67.82	66.75	65.66	65.96	69.13	66.17	69.00	69.62	67.35	**71.13**
Defocus Blur	72.06	76.65	78.78	74.31	75.00	76.46	78.78	77.29	79.78	77.14	77.79	**80.36**
Snow	71.04	72.64	76.39	74.66	74.16	73.25	77.31	74.05	78.50	77.20	74.94	**79.53**
JPEG Compression	71.84	74.38	78.73	75.56	74.77	76.77	80.81	74.61	79.79	77.98	77.94	**82.06**
V20-C Average	69.33	71.59	74.90	72.58	71.99	72.66	75.97	72.41	76.18	75.08	73.89	**77.58**

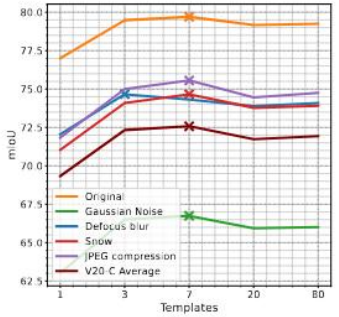


Figure 4: mIoU performance of our method for different numbers of templates.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Test-Time Multi-Prompt Adaptation for Open-Vocabulary Remote Sensing Image Segmentation

Ting Yang[1], Qilong Wang[1]*, Qibin Hou[2], Qinghua Hu[1]
[1]Tianjin University [2]Nankai University
{yting_123, qlwang, huqinghua}@tju.edu.cn, houqb@nankai.edu.cn

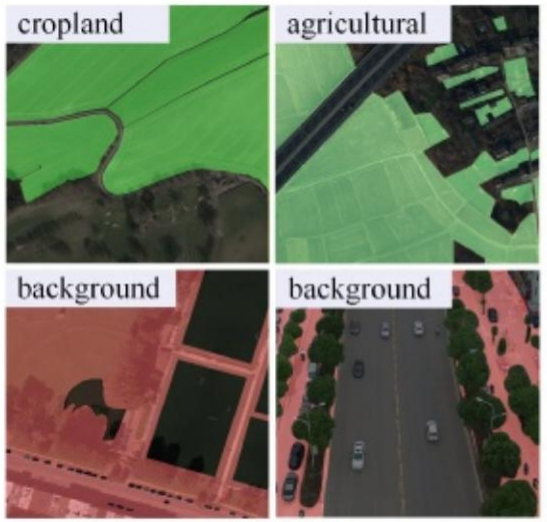
汇报人：蒋明忠

时间：2026.06

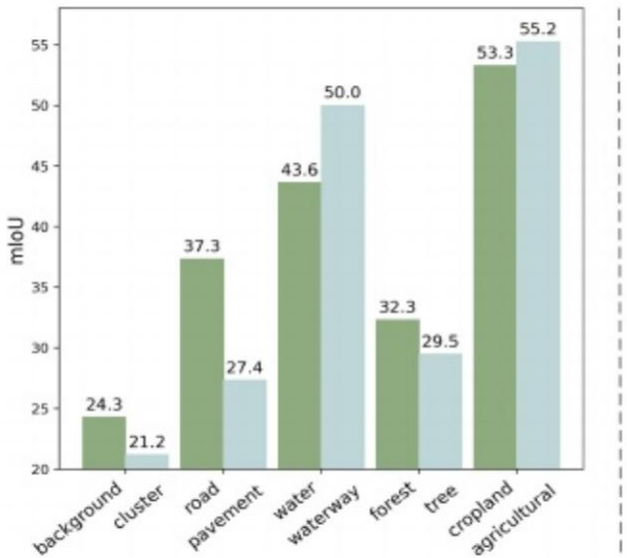
CVPR

Background

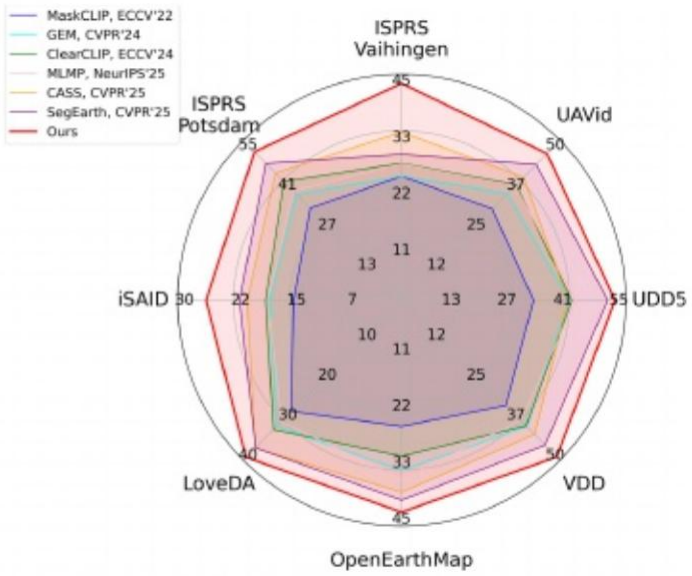
现有方法大多基于 CLIP 这类视觉-语言模型，把图像特征和文本类别特征投到同一个语义空间中，然后计算相似度完成像素分类。问题在于，以前的 OVRSIS 方法主要改进视觉侧特征，比如增强遥感图像的空间细节、尺度鲁棒性、旋转不变性等，却忽略了 文本侧 prompt 本身的歧义。



(a) Textual Ambiguity



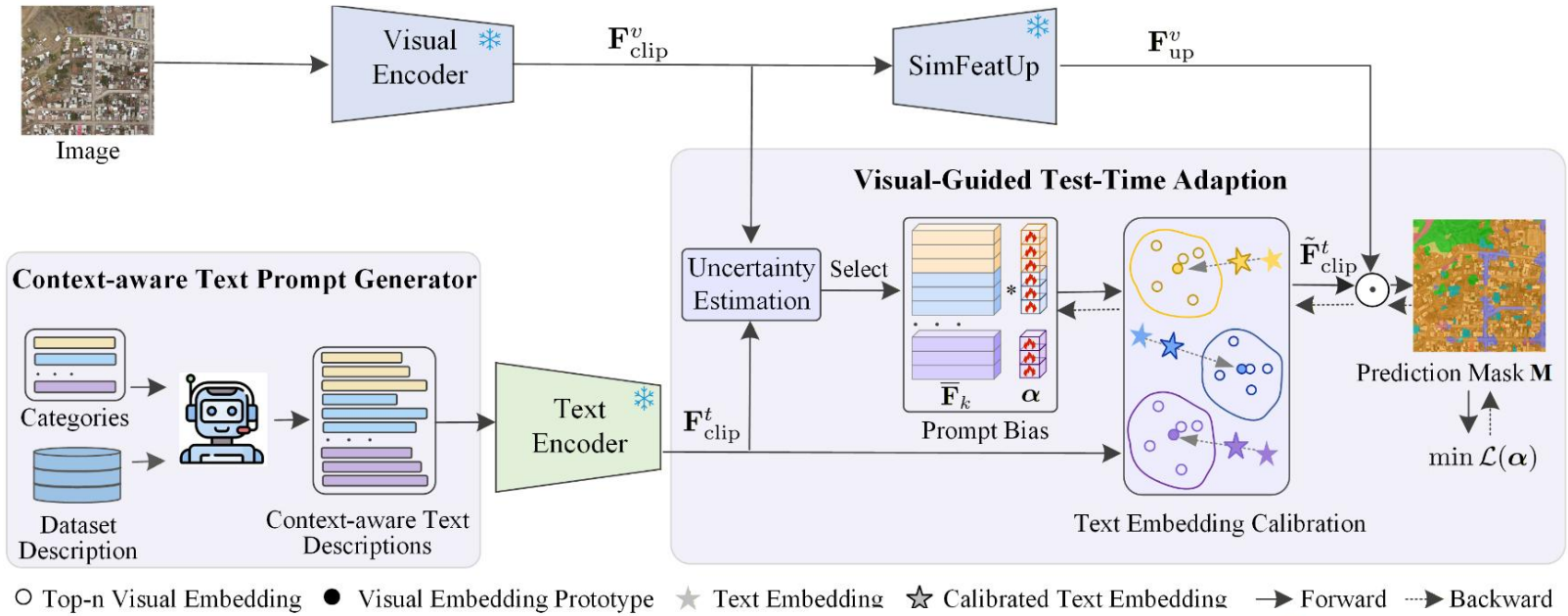
(b) Sensitivity to Text Prompts



(c) Performance Comparison (mIoU, %)

Method

我们的TMPA提出了一个上下文感知文本提示生成器和一个视觉引导测试时间适应 (VGTA)策略。感知文本提示生成一组不同的、上下文感知的描述，以减轻文本歧义，而VGTA在测试时动态地改进生成的提示的嵌入，保证更好的泛化。



$$\tilde{\mathbf{F}}^t_{\text{clip}}(\boldsymbol{\alpha}) = (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{F}^t_{\text{clip}} + \boldsymbol{\alpha} \odot \overline{\mathbf{F}},$$

$$\min \mathcal{L}(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\alpha}^\star} \frac{1}{H'W'} \sum_{x,y} \mathbf{P}^\top_{x,y}(\boldsymbol{\alpha}) \log \mathbf{P}_{x,y}(\boldsymbol{\alpha}),$$

$$\mathbf{P}(\boldsymbol{\alpha}) = \text{softmax}\left(\mathbf{F}^v_{\text{up}} \left(\tilde{\mathbf{F}}^t_{\text{clip}}(\boldsymbol{\alpha})\right)^\top\right), \qquad (8)$$

Experiments

Methods	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid	UDD5	VDD	Avg
MaskCLIP [65]	25.1	27.8	14.5	31.7	24.7	28.6	32.4	32.9	27.2
SCLIP [55]	29.3	30.4	16.1	36.6	28.4	31.4	38.7	37.9	31.1
GEM [4]	33.9	31.6	17.7	36.5	24.7	33.4	41.2	39.5	32.3
ClearCLIP [29]	31.0	32.4	18.2	40.9	27.3	36.2	41.8	39.3	33.4
CASS [27]	38.2	37.0	20.7	43.8	33.5	38.5	40.9	42.0	37.4
MLMP [44]	35.5	30.4	17.9	37.6	27.3	35.9	42.9	37.56	33.1
SegEarth-OV [31]	39.8	36.9	21.7	47.1	29.1	42.5	50.6	45.3	39.1
OVRS* [7]	-	-	93.3	19.9	20.8	-	-	-	-
RSKT-Seg* [30]	-	28.1	93.2	20.3	17.5	17.2	13.9	25.3	-
TMPA (Ours)	**42.2** ↑2.4	**39.7** ↑2.8	**26.2** ↑4.5	**51.1** ↑4.0	**43.4** ↑14.3	**45.9** ↑3.4	**52.4** ↑1.8	**49.0** ↑3.7	**43.7** ↑4.6

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

CoCo-SAM3: Harnessing Concept Conflict in Open-Vocabulary Semantic Segmentation

Yanhui Chen[1], Baoyao Yang[1], Siqi Liu[2], and Jingchao Wang[3]

[1] School of Computers, Guangdong University of Technology [2] Shenzhen Research Institute of Big Data [3] Peking University

chenyanhui91@mails.gdut.edu.cn, ybaoyao@gdut.edu.cn, siqiliu@sribd.cn, ethanwangjc@163.com

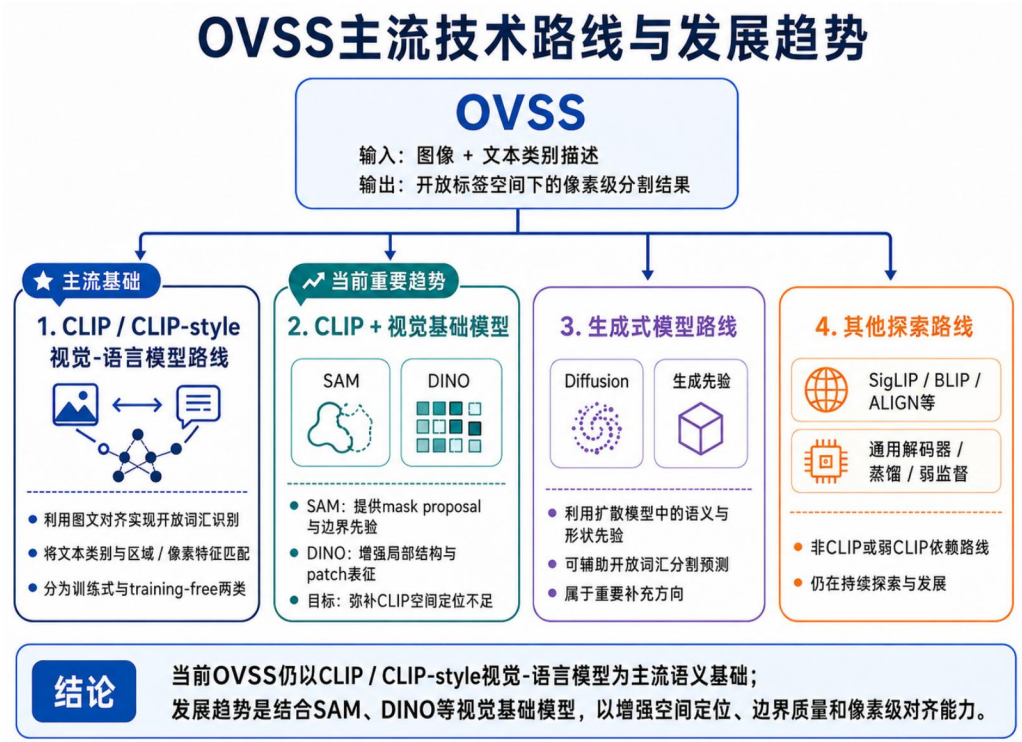
汇报人：蒋明忠

时间：2026.06

Background

开放词汇语义分割（Open-Vocabulary Semantic Segmentation, OVSS）旨在根据自然语言描述的类别概念，在开放标签空间中实现像素级语义分割。与传统语义分割方法依赖固定标签集、且要求训练类别与测试类别保持一致不同，OVSS允许在推理阶段动态指定类别集合，并要求模型能够泛化到未见过的类别以及多样化的文本表达方式。

Train-free OVSS方法则进一步避免针对分割任务进行额外训练或微调，主要利用CLIP等预训练视觉-语言模型已有的图文对齐能力来获得开放词汇识别能力。由于CLIP原本主要面向图像级图文对齐，直接用于语义分割时往往存在局部感知不足、空间定位不精细和类别响应噪声等问题。



Background

作为新一代提示性分割基础模型，SAM3将OVSS从相似性驱动的判别预测进一步推进到概念条件化范式：以一个概念提示为条件，直接生成对应概念的分割掩码，提供了一种更为准确的分割方法。

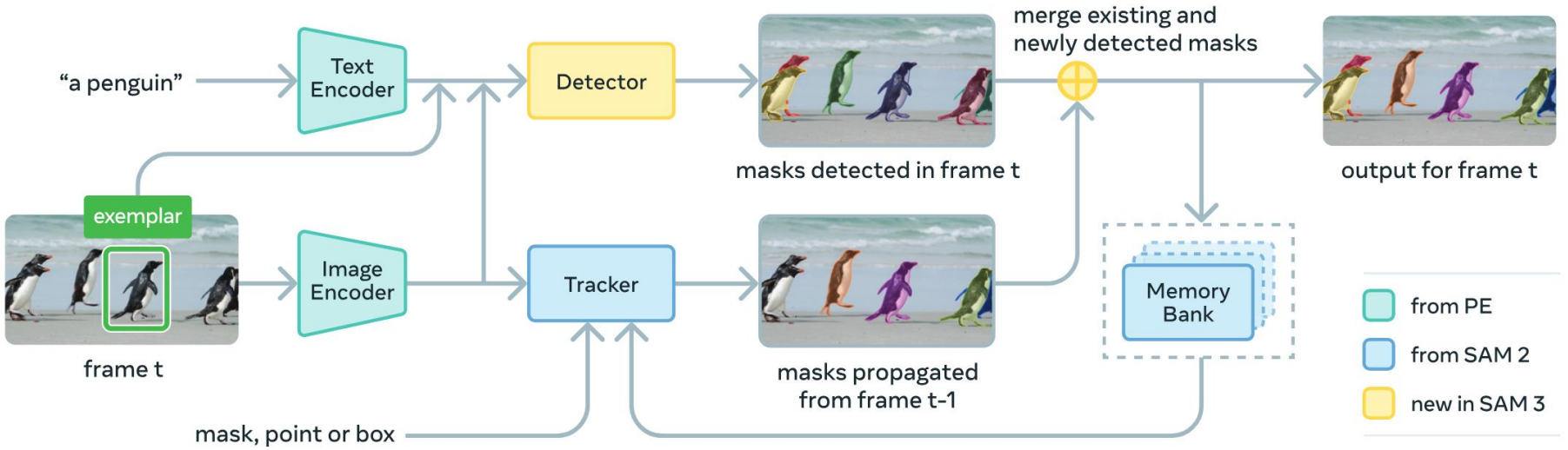
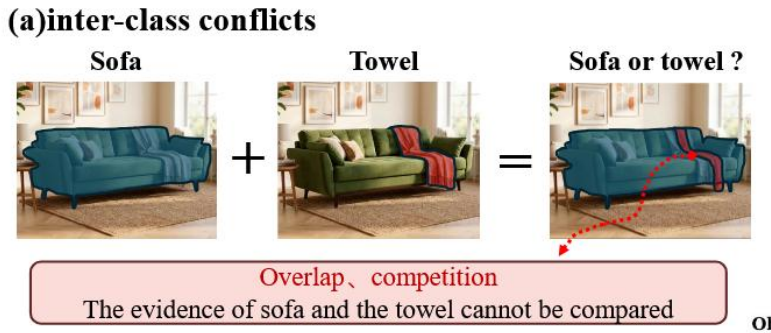


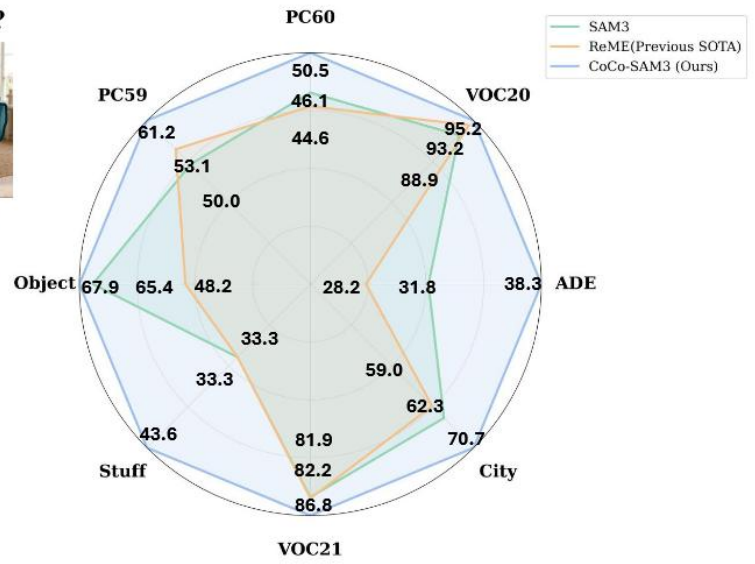
Figure 4 SAM 3 architecture overview. See Fig. 10 for a more detailed diagram.

Background

SAM3通过引入提示驱动的掩码生成范式来推进开放词汇语义分割。

然而，在多类开放词汇场景中,由不同类别提示独立生成的掩码缺乏统一的、 类间可比较的证据尺度,往往导致覆盖重叠和竞争不稳定。此外，同一概念的同义表达往往会激活不一致的语义和空间证据，导致类内漂移，加剧类间冲突, 损害整体推理稳定性。





Method

从多类开放词汇联合推理的角度出发，我们发现了SAM3的两个关键局限性：缺乏跨概念竞争的统一证据尺度，以及开放命名多样性下缺乏类内语义一致性维护。我们将导致的系统性不稳定归因于SAM3的输出组织及其即时条件反应机制。

为了解决这些问题，我们提出了CoCo-SAM3(概念冲突 SAM3),它明确地将推理解耦为类内增强和类间竞争。我们的方法首先对同义提示的证据进行对齐和聚合,以加强概念一致性。然后,它在统一的可比较尺度上执行类间竞争,实现所有候选类之间的直接逐像素比较。

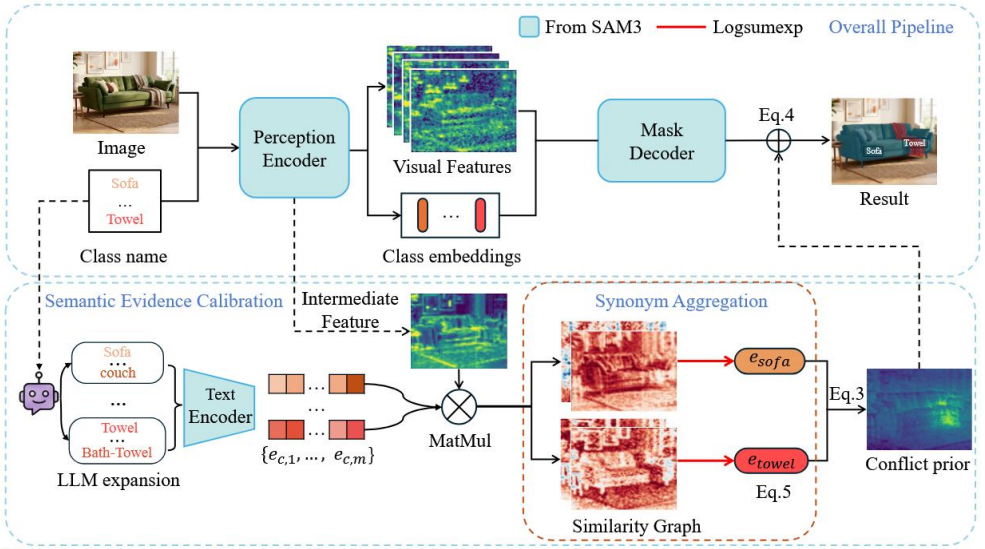


Fig. 2: The overview of **CoCo-SAM3**. We enhance semantic-evidence consistency via intra-class synonym aggregation and build a unified-scale conflict prior to stabilize inter-class competition, yielding stable open-vocabulary semantic segmentation.

Method

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

语义证据校准(Semantic Evidence Calibration, SEC)：通过将文本概念与来自感知编码器的中间层视觉表示进行匹配，我们明确地为像素级语义结构一致性建模,为多类联合推理提供稳定的语义先验,并缓解由独立生成的掩码引起的冲突。

SAM3 对每个 prompt 输出一个 mask probability：

$$P_c^{sam}(x)$$

这里 c 是类别，x 是像素位置。

但这个概率本身不适合直接跨类别比较。所以作者额外构造一个 semantic prior。

他们从 SAM3 的 Perception Encoder 中间层取 dense feature：

$$f(x)$$

对每个类别文本 embedding：

$$e_c$$

做 L2 normalize 后计算余弦相似度：

$$u_c(x) = e_c^\top f(x)$$

这个 $u_c(x)$ 表示像素 x 和类别 c 的语义匹配强度。然后对所有候选类别做 softmax：

$$\pi_c(x) = \frac{\exp(u_c(x))}{\sum_{c'} \exp(u_{c'}(x))}$$

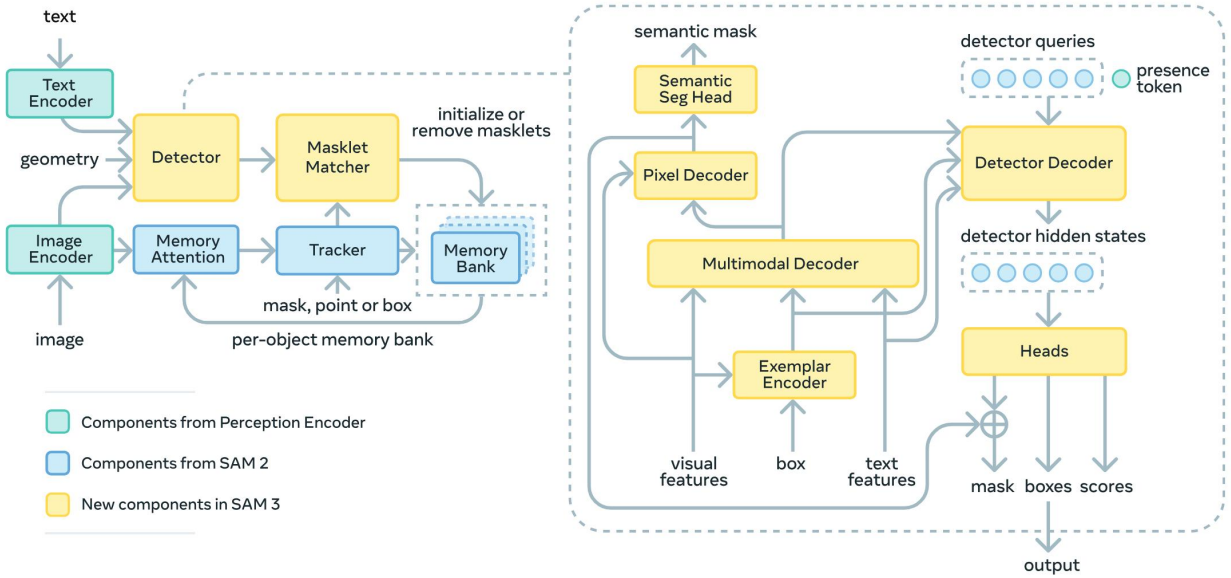


Figure 10 SAM 3 architecture. New components are in yellow, SAM 2 (Ravi et al., 2024) in blue and PE (Bolya et al., 2025) in cyan.

Method

语义证据校准(Semantic Evidence Calibration, SEC)：通过将文本概念与来自感知编码器的中间层视觉表示进行匹配，我们明确地为像素级语义结构一致性建模,为多类联合推理提供稳定的语义先验,并缓解由独立生成的掩码引起的冲突。

SAM3 对每个 prompt 输出一个 mask probability：

$$P_c^{sam}(x)$$

这里 c 是类别，x 是像素位置。

但这个概率本身不适合直接跨类别比较。所以作者额外构造一个 semantic prior。

他们从 SAM3 的 Perception Encoder 中间层取 dense feature：

$$f(x)$$

对每个类别文本 embedding：

$$e_c$$

做 L2 normalize 后计算余弦相似度：

$$u_c(x) = e_c^\top f(x)$$

这个 $u_c(x)$ 表示像素 x 和类别 c 的语义匹配强度。然后对所有候选类别做 softmax：

$$\pi_c(x) = \frac{\exp(u_c(x))}{\sum_{c'} \exp(u_{c'}(x))}$$

然后作者把 SAM3 的 mask probability 转成 logit：

$$\log \frac{P_c^{sam}(x)}{1 - P_c^{sam}(x)}$$

再把 semantic prior 的 log 加进去：

$$S_c(x) = \log \frac{P_c^{sam}(x)}{1 - P_c^{sam}(x)} + \lambda_{prior} \log \pi_c(x) + z_c$$

其中：

- 第一项：SAM3 mask 的结构证据，主要负责边界和区域；
- 第二项：semantic prior，负责跨类别竞争；
- 第三项：SAM3 presence logit，负责图像级类别存在性偏置。

最后：

$$\hat{y}(x) = \arg\max_c S_c(x)$$

Method

同义词聚合(Synonym Aggregation，SA)：在多类开放词汇推理中,能否稳定抑制类间冲突往往取决于类内证据能否形成 稳定的代表性信号：如果同一概念在不同的语言实现下激活了实质上不同的语义证据,其在类间竞争中的相对优势就会随着措辞的变化而波动,从而放大覆盖和标签的不稳定性

对于每个类别 c，作者用 LLM 生成一组同义表达：

$$\mathcal{S}_c = \{s_{c,1}, ..., s_{c,m}\}$$

例如：

```
sofa → sofa, couch
towel → towel, bath towel
```

每个 synonym prompt 都编码成文本 embedding：

$$e_{c,j}$$

然后分别和 dense feature 做相似度：

$$u_{c,j}(x) = e_{c,j}^{\top} f(x)$$

接着用 LogSumExp 聚合：

$$\tilde{u}_c(x) = \log \sum_j \exp\left(\frac{e_{c,j}^{\top} f(x)}{\tau_s}\right)$$

为什么用 LogSumExp?

因为它介于 average 和 max 之间：

- 比 average 更能突出强匹配 synonym；
- 比 max 更能保留多个 synonym 的互补证据；
- 当多个 synonym 都支持同一位置时，证据会累积；
- 当某个 synonym 噪声大时，不会像平均那样严重拖累。

$$u_c(x) = e_c^{\top} f(x)$$

Experiments

Method	*with background*			*without background*					Avg.
	V21	PC60	COCO-O	V20	PC59	COCO-S	City	ADE	
Training-based									
GroupViT [39] (CVPR'22)	50.4	18.7	27.5	79.7	23.4	15.3	11.1	9.2	29.4
TCL [31] (CVPR'23)	51.2	24.3	30.4	77.5	30.3	19.6	23.1	14.9	33.9
CLIP-DINOiser [36] (ECCV'24)	62.1	32.4	34.8	80.9	35.9	24.6	31.7	20.0	40.3
Talk2DINO [2] (ICCV'25)	65.8	37.7	45.1	88.5	42.4	30.2	38.1	22.5	46.3
Training-free CLIP-Only									
CLIP [29] (ICML'21)	18.6	7.8	6.5	49.1	11.2	7.2	6.7	3.2	13.8
CaR [15] (CVPR'24)	48.6	13.6	15.4	73.7	18.4	–	–	5.4	–
CLIPtrase [32] (ECCV'24)	50.9	29.9	43.6	81.0	33.8	22.8	–	16.4	–
ClearCLIP [19] (ECCV'24)	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	38.1
SCLIP [35] (ECCV'24)	59.1	30.4	30.5	80.4	34.1	22.4	32.2	16.1	38.2
NACLIP [14] (WACV'25)	58.9	32.2	33.2	79.7	35.2	23.3	35.5	17.4	39.4
SFP [16] (ICCV'25)	63.9	37.2	37.9	84.5	39.9	26.4	41.1	20.8	44.0
RF-CLIP [22] (AAAI'26)	67.2	37.9	39.1	87.0	41.4	27.5	43.0	21.0	45.5
Training-free CLIP-VFM									
FreeDA [1] (CVPR'24)	51.8	35.3	36.3	84.3	39.7	25.7	34.1	20.8	41.0
ProxyCLIP [20] (ECCV'24)	58.6	33.8	37.4	83.0	37.2	25.4	33.9	19.7	41.1
CASS [17] (CVPR'25)	65.8	36.7	37.8	87.8	40.2	26.7	39.4	20.4	44.4
CorrCLIP [42] (ICCV'25)	76.7	44.9	49.4	91.5	50.8	34.0	51.1	30.7	53.6
Trident [34] (ICCV'25)	67.1	38.6	41.1	84.5	42.2	28.3	42.9	21.9	45.8
ReME [41] (ICCV'25)	82.2	44.6	48.2	93.2	53.1	33.3	59.0	28.2	55.2
Training-free SAM3									
SAM3 [7] (ICLR'26)	81.9	46.1	65.4	88.9	50.0	33.3	62.3	31.8	57.5
CoCo-SAM3 (Ours)	86.8	50.5	67.9	95.2	61.2	43.6	70.7	38.3	64.3

Experiments

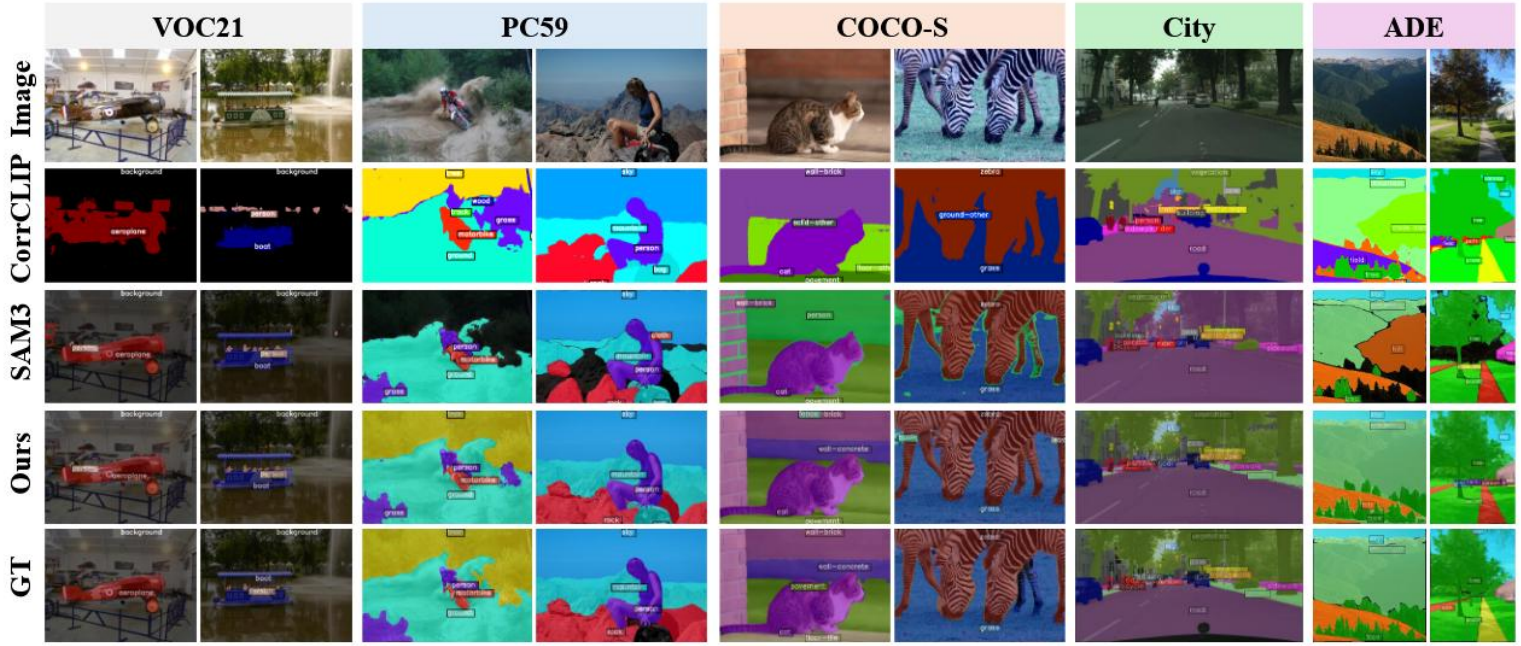


Fig. 3: Qualitative comparisons. "GT" denotes the ground truth.

SEC	SA	VOC20	PC59	COCO-S	City	ADE	Avg.
×	×	88.9	50.0	33.3	62.3	31.8	53.3
✓	×	94.8	59.1	43.0	67.9	37.3	60.4
✓	✓	**95.2**	**61.2**	**43.6**	**70.7**	**38.3**	**61.8**

Thanks