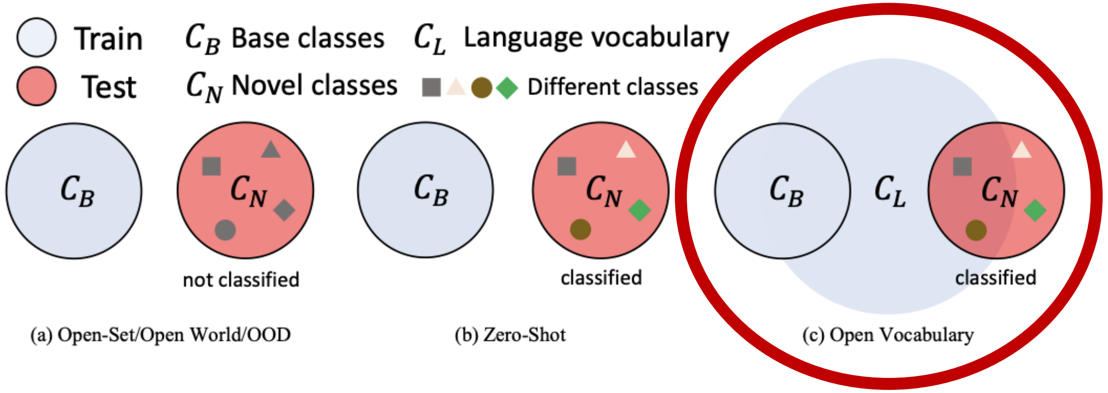
Uncovering Prototypical Knowledge for Weakly Open-Vocabulary Semantic Segmentation

Fei Zhang, Tianfei Zhou, Boyang Li, Hao He,
Chaofan Ma, Tianjiao Zhang, Jiangchao Yao, Ya Zhang, Yanfeng Wang

Weakly Open-Vocabulary Semantic Segmentation

➢ OVSS targets on the *Segmentor* that can segment the **novel class** from **Large Language Knowledge**.



Saturn, Blossom

Golden Gate, Yacht

➢ WOVSS focuses on training a OV *Segmentor* with only image-text pairs.

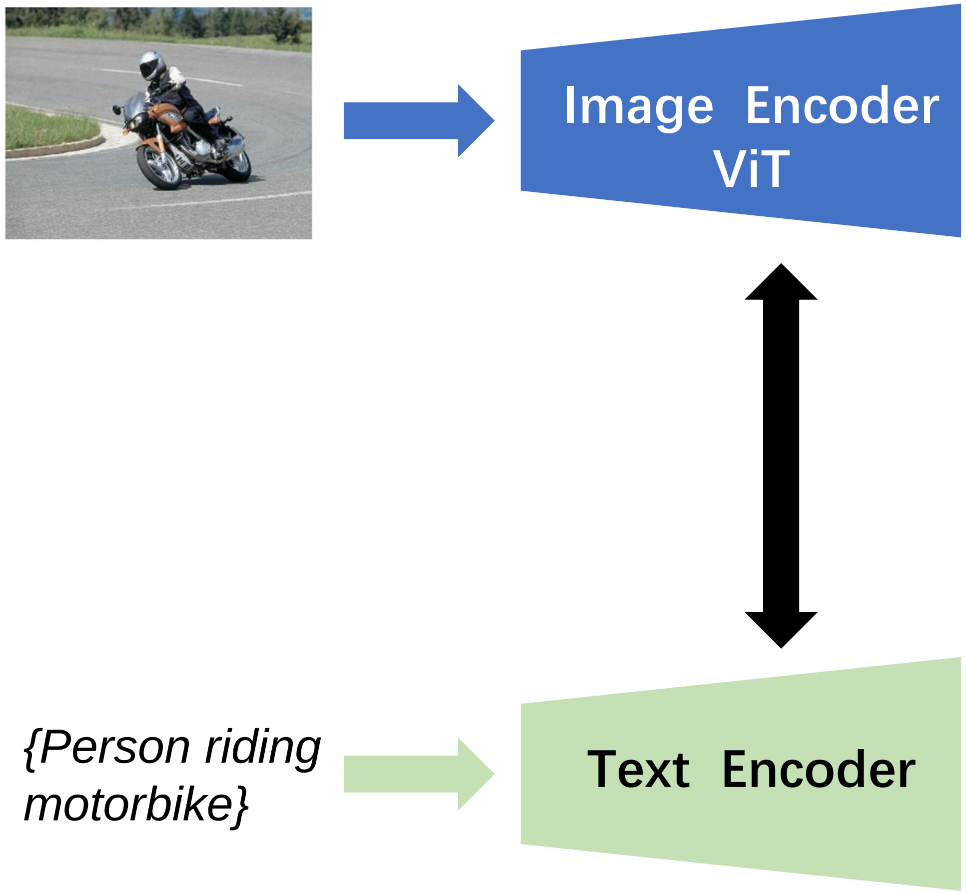
Expensive!

⊕ *{Person riding motorbike}*

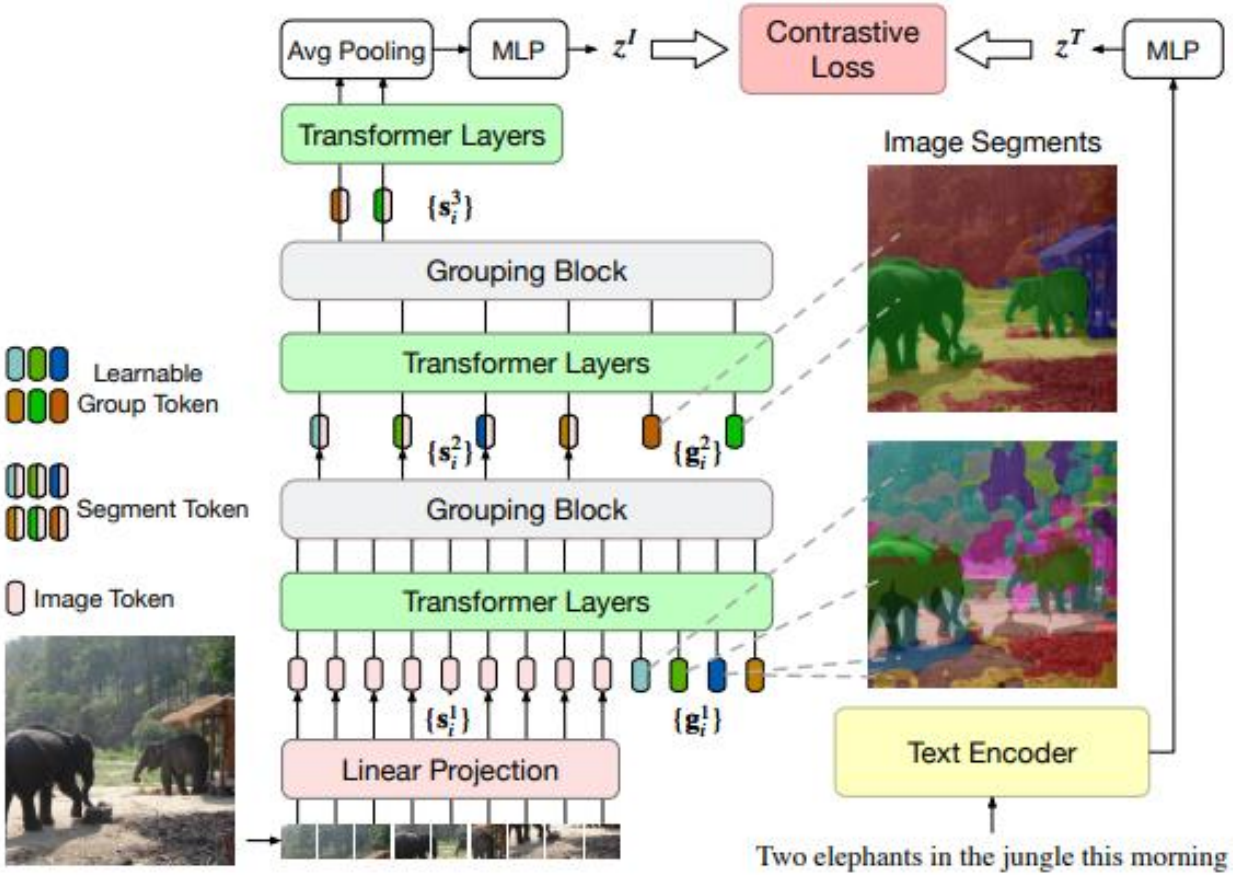
→ **Segmentor**

How to address WOVSS?

- *Image-Text Alignment* is the baseline.



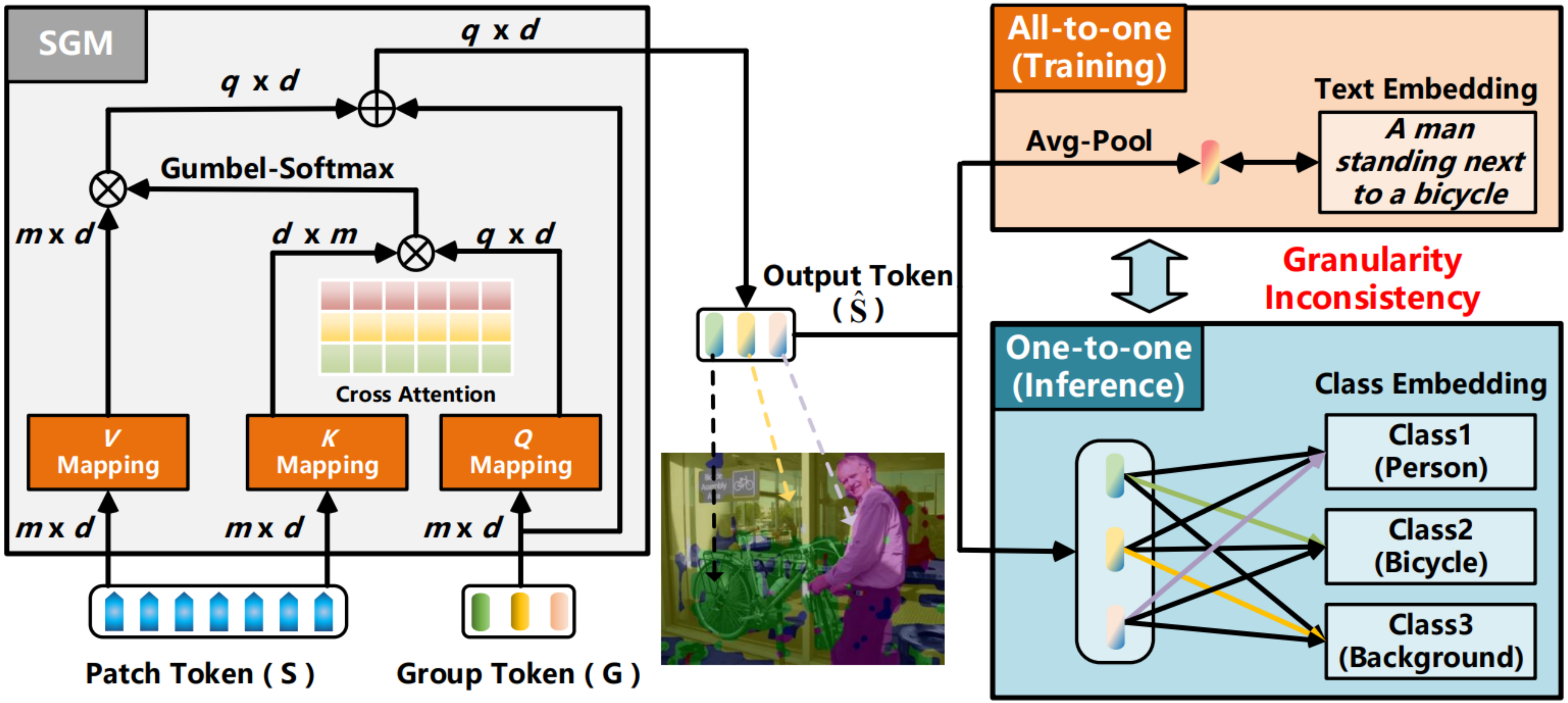
- *Semantic Grouping Module* (SGM) enables **segmenting ability**.



GroupViT, CVPR’22

Granularity Inconsistency in SGM

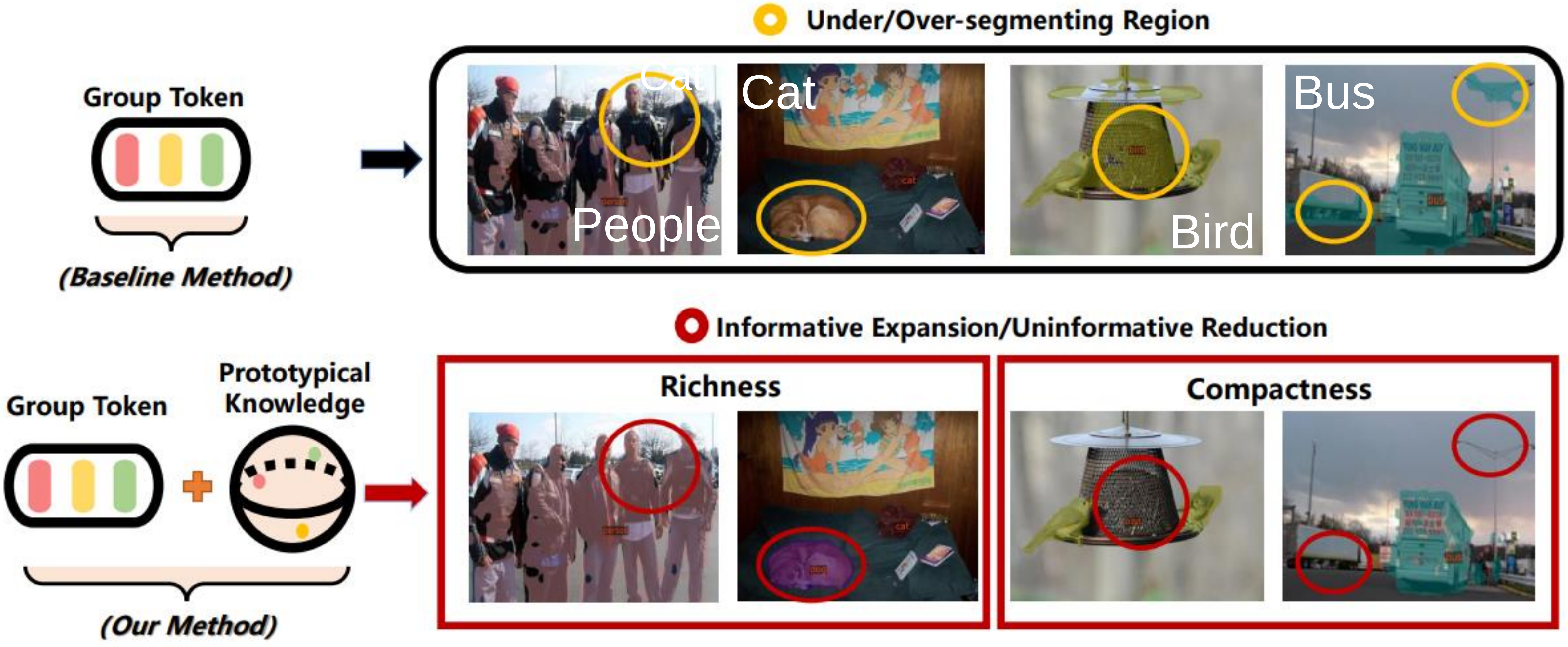
- Granularity Inconsistency: all-to-one (training) *vs* one-to-one (inference).



Group tokens/centroids lack *explicit supervision* **?**

Finding the proper supervision

➢ *Prototypical knowledge -> Compactness* and *Richness*.



Non-learnable Prototypical Regularization

NEURAL INFORMATION PROCESSING SYSTEMS

- NPR -> *Generating* the supervision and *Regularizing* the group token.

$$y_{ij} = \frac{\mathcal{Z}(\boldsymbol{v}_j|\boldsymbol{p}_i)}{\sum_{i=1}^{q} \mathcal{Z}(\boldsymbol{v}_j|\boldsymbol{p}_i)} = \frac{\exp(\boldsymbol{p}_i \boldsymbol{v}_j^\top)}{\sum_{i=1}^{q} \exp(\boldsymbol{p}_i \boldsymbol{v}_j^\top)} \qquad \boldsymbol{p}_i = \frac{\sum_{j=1}^{m} y_{ij} \boldsymbol{v}_j}{\sum_{j=1}^{m} y_{ij}}$$

Algorithm 1 Non-learnable Prototypical Regularization (NPR)

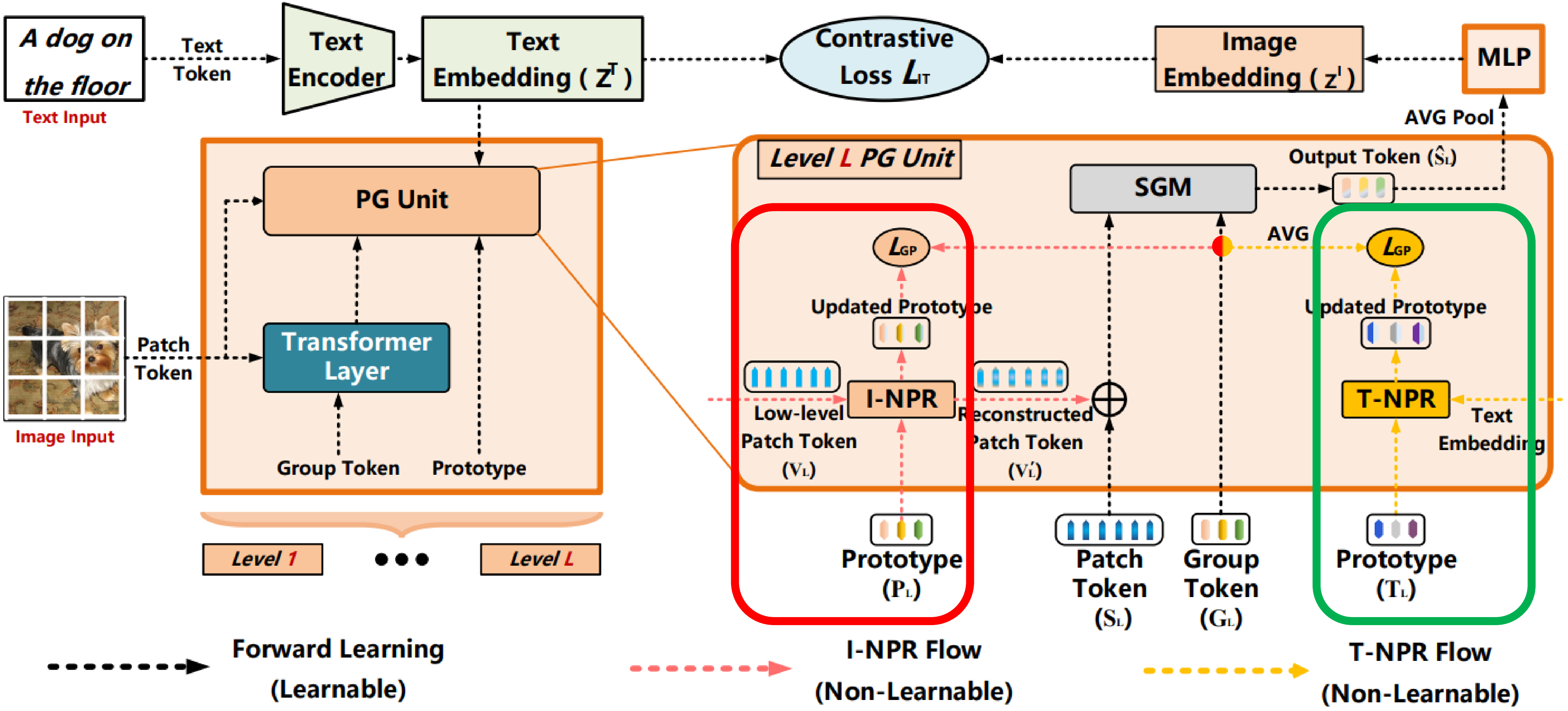
Require: Group tokens $\boldsymbol{G} \in \mathbb{R}^{q \times d}$, prototypes $\boldsymbol{P} \in \mathbb{R}^{q \times d}$, prototypical source features $\boldsymbol{V} \in \mathbb{R}^{m \times d}$, iterations T ($T = 10$ in our setting), selecting threshold ϕ.

1: ▷ **Prototype Generation**
2: **for** iteration $t = 1$ to T **do**
3: (**E-step**) **Calculate** the probability of $\boldsymbol{V}$ belonging to $\boldsymbol{P}$ in Eq. (1)
4: (**M-step**) Update the prototypes $\boldsymbol{P}$ by using Eq. (2)
5: **end for**
6: ▷ **Prototype Supervision**
7: **Generate** the matched prototypes $\boldsymbol{P}^{\mathrm{h}}$ by using the Hungarian matching between $\boldsymbol{P}$ and $\boldsymbol{G}$
8: **Select** the matched pairs ($\boldsymbol{P}^{\mathrm{h}}$, $\boldsymbol{G}$) whose similarity scores are beyond ϕ
9: **Regularize** the selected $\boldsymbol{G}$ with the matched $\boldsymbol{P}^{\mathrm{h}}$ by using $\mathcal{L}_{\mathrm{PG}}$ in Eq. (3)

$$\mathcal{L}_{\mathrm{PG}}(\boldsymbol{G}, \boldsymbol{P}^{\mathrm{h}}) = -\frac{1}{q} \sum_{i=1}^{q} \left(\log \frac{\exp(\mathcal{S}(\boldsymbol{g}_i, \boldsymbol{p}_i^{\mathrm{h}})/\tau)}{\sum_{j=1}^{q} \exp(\mathcal{S}(\boldsymbol{g}_i, \boldsymbol{p}_j^{\mathrm{h}})/\tau)} + \log \frac{\exp(\mathcal{S}(\boldsymbol{p}_i^{\mathrm{h}}, \boldsymbol{g}_i)/\tau)}{\sum_{j=1}^{q} \exp(\mathcal{S}(\boldsymbol{p}_i^{\mathrm{h}}, \boldsymbol{g}_j)/\tau)}\right)$$

Prototypical Guidance Segmentation Network

➢ PGSeg -> Instantiate NPR with multi-modal prototypical knowledge.



Experiments

- Backbone: ViT-S
- Training Dataset: CC12M/RedCaps12M
- Evaluation Dataset: VOC12/Context/COCO

Methods	Training Data (volume)	Pre-trained Models	VOC12	Context	COCO
RECO [45]	CC400M [39] + ImageNet1M (401M)	CLIP [39] + MOCO [21]	25.1	19.9	15.7
MaskCLIP [55]	CC400M [39] (400M)	CLIP [39]	29.3	21.1	15.5
ViL-Seg [33]	CC12M [7] (12M)	✘	34.4	16.3	16.4
MaskCLIP [55]	CC400M [39] + *ST* (400M)	CLIP [39]	38.8	**23.6**	20.6
GroupViT [50]	CC12M [7] (12M)	✘	41.1	18.2	18.4
OVSegmentor [51]	CC12M [7] + ImageNet1M [11] (13M)	BERT [13] + DINO [5]	44.5	18.3	19.0
PGSeg	CC12M [7] (12M)	✘	**<u>49.0</u>**	<u>20.6</u>	**<u>22.9</u>**
GroupViT [50]	CC12M [7] + RedCaps12M [12] (24M)	✘	50.8	23.6	27.5
SegCLIP [35]	CC403M [39, 7] + COCO400k [32] (403.4M)	CLIP [39]	52.6	**24.7**	26.5
GroupViT [50]	CC12M [7] + YFCC14M [46] (26M)	✘	52.3	22.4	20.9
ViewCO [41]	CC12M [7] + YFCC14M [46] (26M)	✘	52.4	23.0	23.5
PGSeg	CC12M [7] + RedCaps12M [12] (24M)	✘	**<u>53.2</u>**	<u>23.8</u>	**<u>28.7</u>**

PGSeg achieves SOTA performance.

Experiments

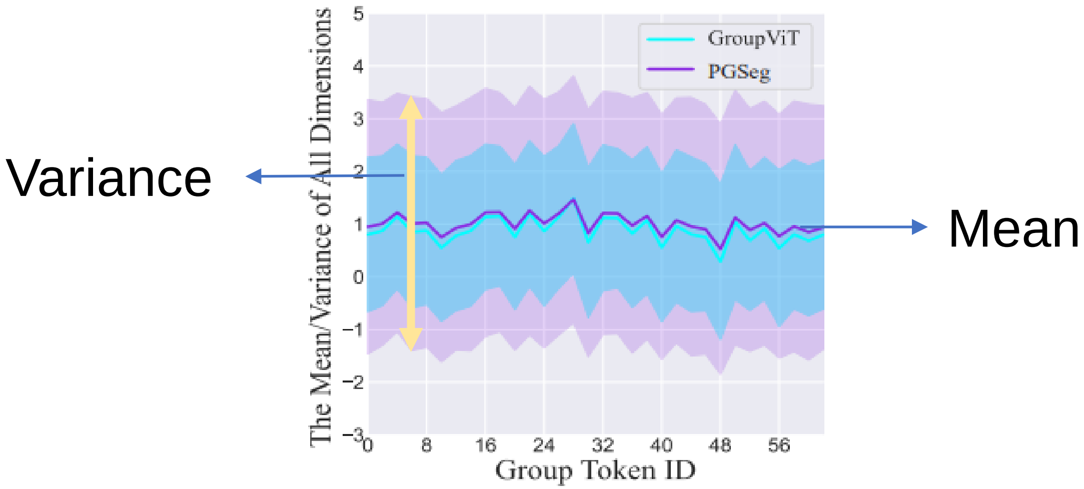
NEURAL INFORMATION PROCESSING SYSTEMS

- Compactness: More *compact* clusters.

GroupViT

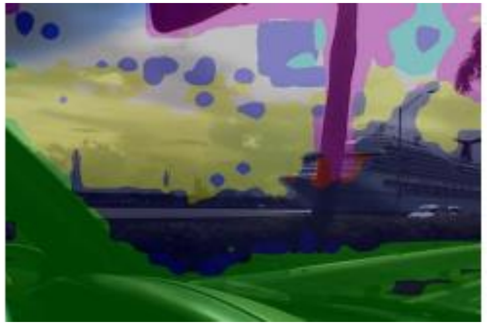
PGSeg

- Richness: *Richer* feature representation.



- Visualized results on PASCAL VOC12.

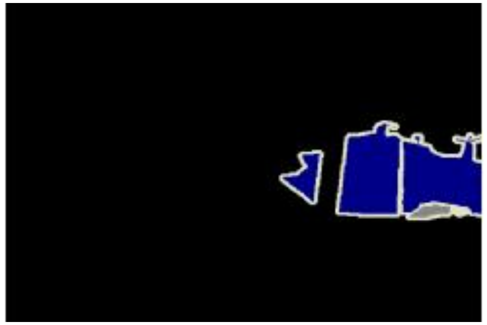
Input



GroupViT

PGSeg

Output



Ground Truth

Experiments

- PGSeg (24 Million Image-text pair) *v.s.* SAM (**11 Billion** Images+**1 Billion** Mask).

Comparable object-level segmentation with only image-text supervision.

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks for Watching!