JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

IFSeg: **I**mage-**F**ree Semantic **Seg**mentation via Vision-Language Model

Sukmin Yun*, Seong Hyeon Park*, Paul Hongsuck Seo, Jinwoo Shin

Poster: TUE-AM-282

GitHub: https://github.com/alinlab/ifseg

Paper: https://arxiv.org/abs/2302.14115

KAIST

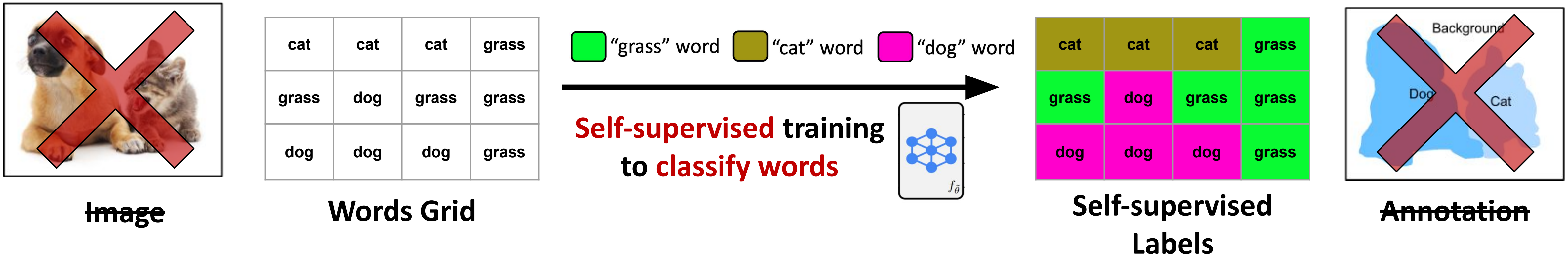
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Google

TL;DR: Image-Free Semantic Segmentation

Image-Free semantic **Seg**mentation (**IFSeg**) via Vision-Language (VL) models

- We propose a novel **self-supervision** method enabling **zero-shot semantic segmentation**
  - Learning to classify **category words** can adapt **pre-trained VL model** for image segmentation!
    no images, no dense annotations are required at all!



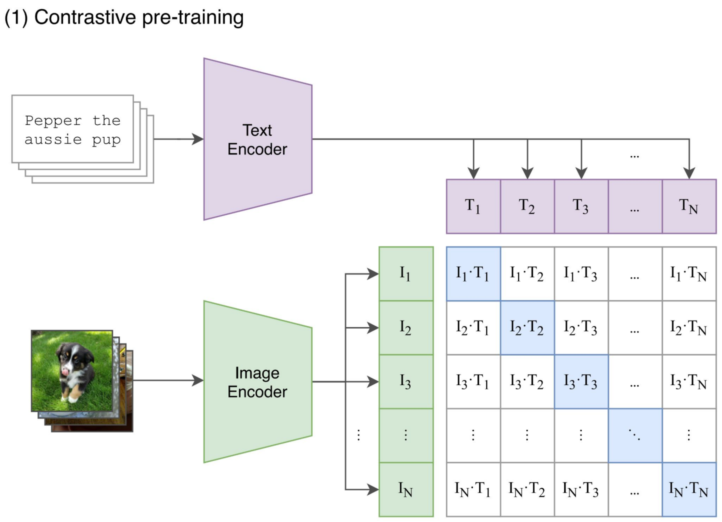
Zero-shot inference on real images

Motivations: VL pre-training for semantic segmentation

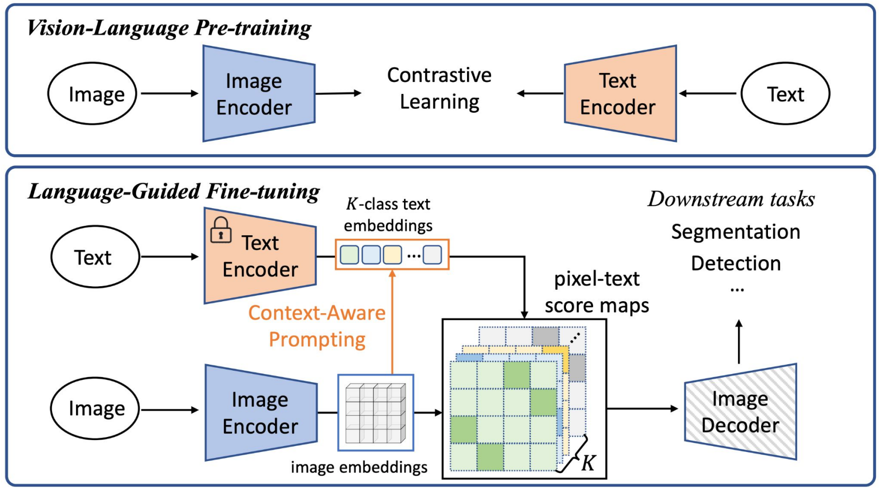
VL pre-training has recently gained attention for its transferability on novel concepts in various visual tasks

- Yet, **VL-driven segmentation** has been under-explored (*e.g.*, image-level vs. pixel-level tasks)
- A *trivial and **expensive*** approach has been fine-tuning with task-specific dataset (**images** and **dense annotation**)

e.g., Contrastive Image Language Pretraining (CLIP)[1]-based segmentation models



CLIP[1] architecture



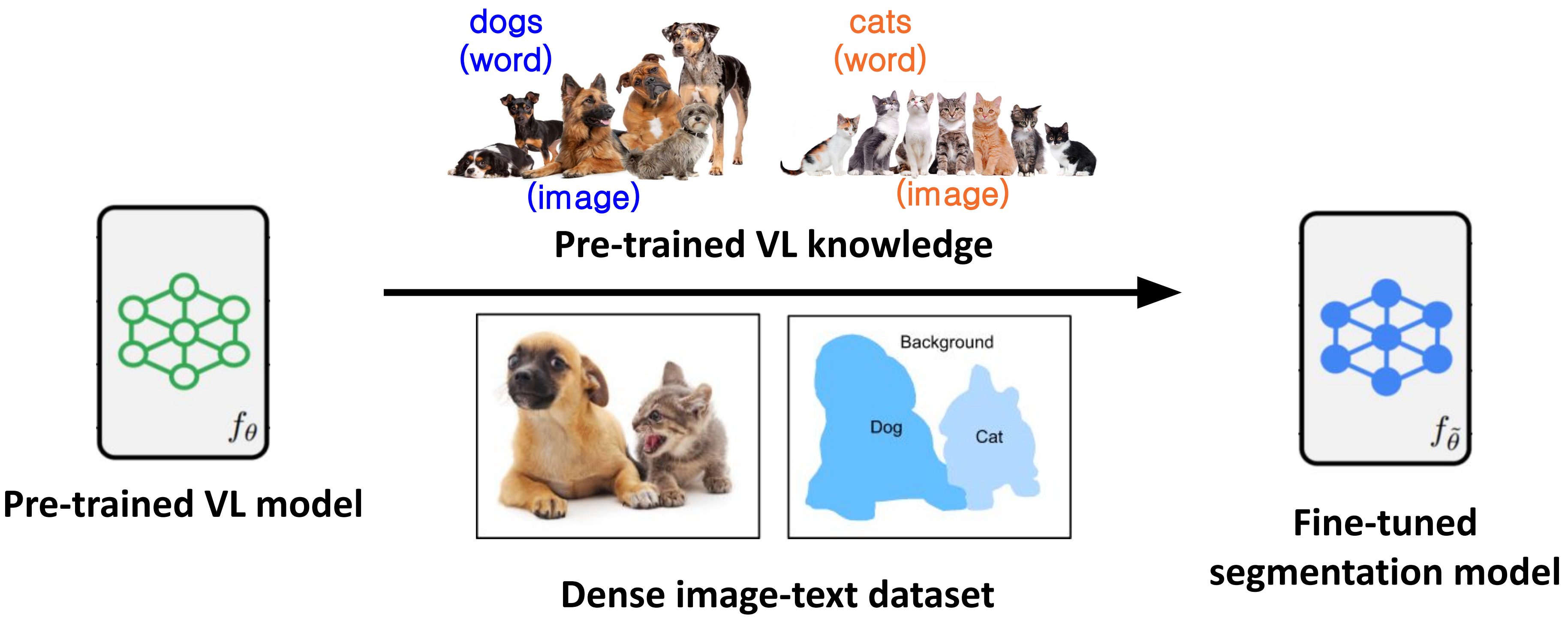
DenseCLIP[2] model

1. Radford et al. Learning transferable visual models from natural language supervision. In ICML 2021. https://arxiv.org/abs/2103.00020
2. Rao et al. Denseclip: Language-guided dense prediction with context-aware prompting. In CVPR 2022. https://arxiv.org/abs/2112.01518

Motivations: VL pre-training for semantic segmentation

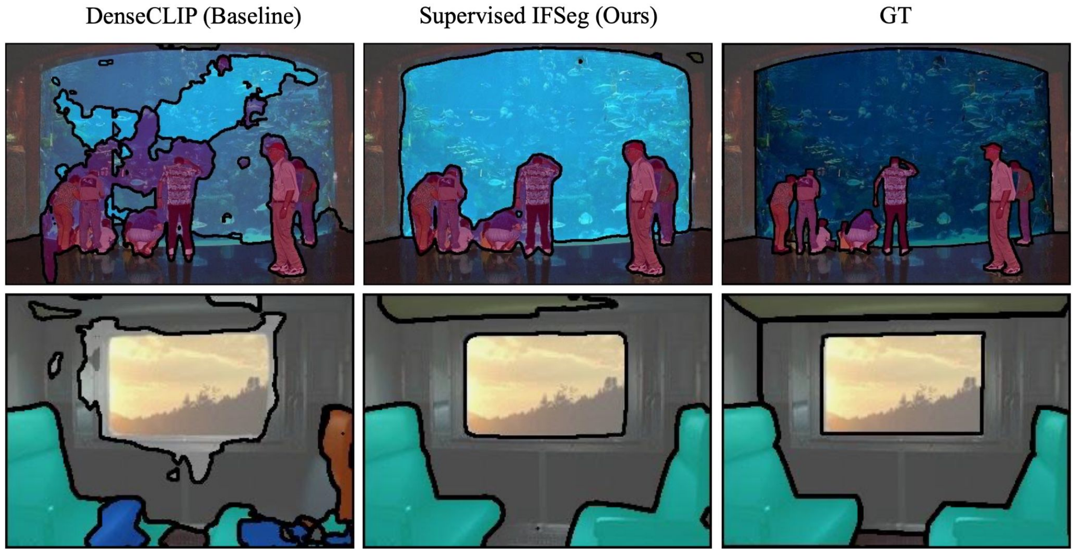
Large-scale VL models tend to have "**open-vocabulary**" (*e.g.*, thousands of classes) knowledge of visual objects

- VL pre-training provides a **good starting point** for **recognizing arbitrary classes** (*e.g.*, the **category word** representation may act as a **zero-shot classifier**)



Motivations: VL pre-training for semantic segmentation

The fine-tuned VL semantic segmentation models show improved performances:



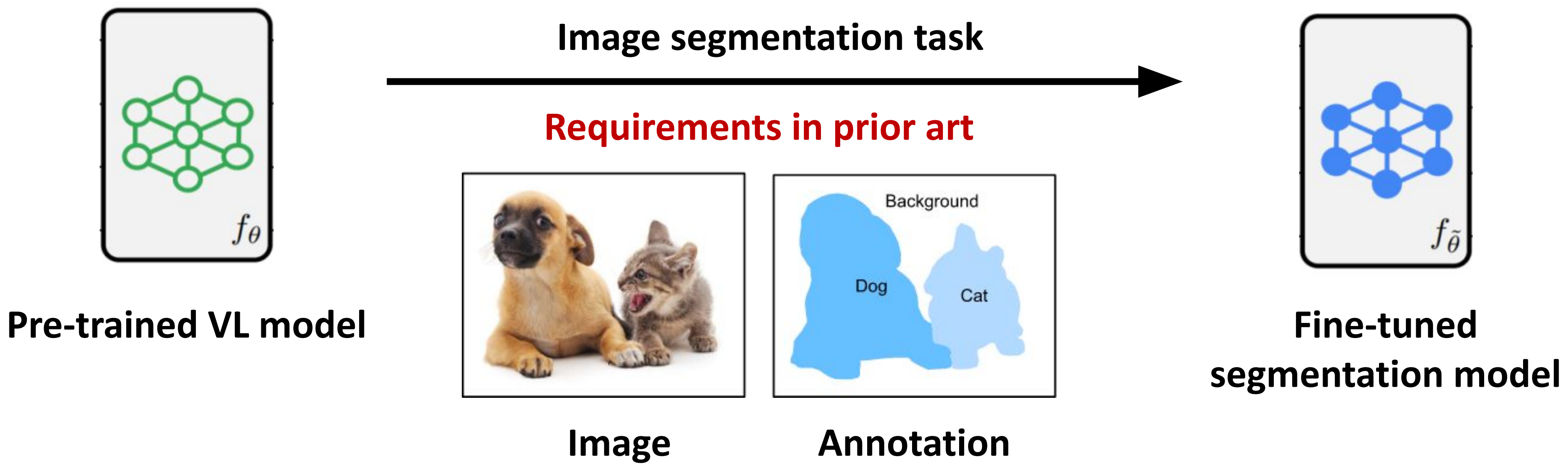
Method	Backbone	mIoU
Semantic FPN [23]	ResNet-101	40.4
UPerNet [46]	ResNet-101	43.8
CLIP + Semantic FPN [32, 34]	ResNet-101	42.7
DenseCLIP + Semantic FPN [34]	ResNet-101	45.1
IFSeg (ours)	ResNet-101	**47.1**

Table 5. **Comparison in supervised semantic segmentation.** We report the mIoU metric evaluated on the 150 semantic cateogires of the ADE20K benchmark. We follow training configurations of DenseCLIP, such as image resolutions and training iterations.

However, is the supervised fine-tuning the best we can leverage pre-trained VL models...?

Motivations: VL pre-training for semantic segmentation

Requirements in fine-tuning: task-specific dataset (**images** and **dense annotation**)

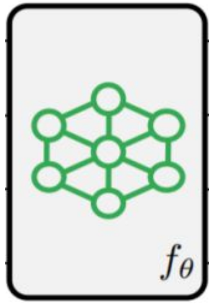


Can we better/fully utilize the "aligned VL representation" for semantic segmentation, possibly without tuning with image data and human-annotated supervision?

Motivations: VL pre-training for semantic segmentation

Idea: If Vision and Language (word) have an aligned representation, words could replace the input images

- Specifically, we may utilize the **semantic category words** to **fully replace** the **images** and **dense annotation**!



Pre-trained VL model

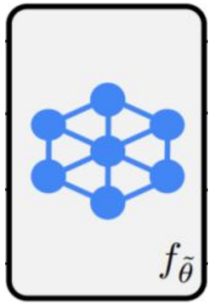
Image segmentation task

~~Requirements in prior art~~

~~Image~~ **~~Annotation~~**

Category Words

***e.g.*, dogs, cats, etc.**

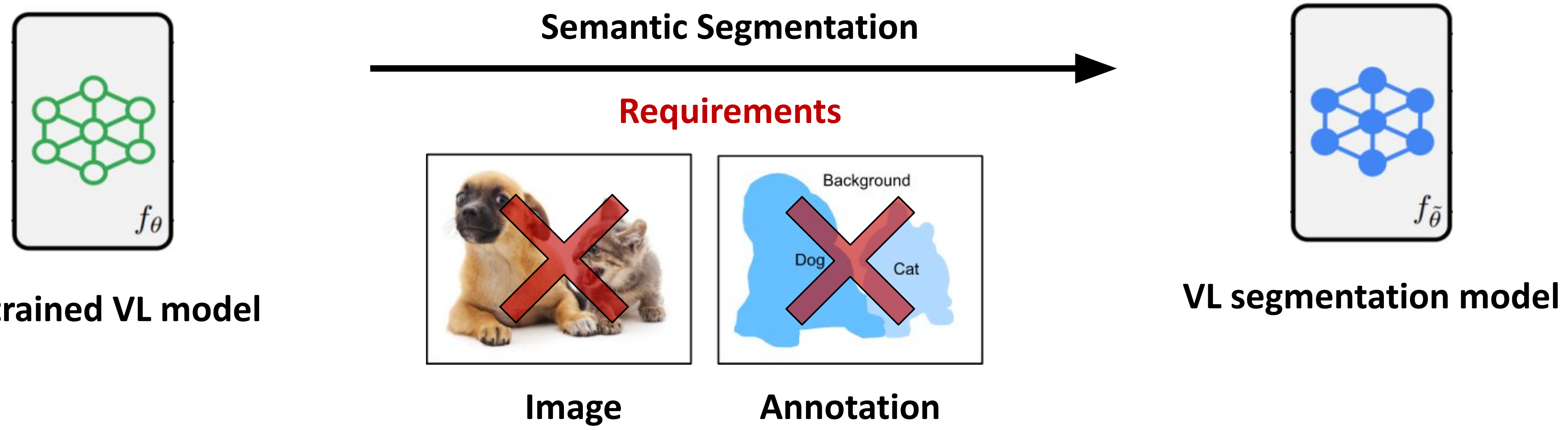


VL segmentation model

Contributions: Image-free Semantic Segmentation Task

We introduce a novel **image-free semantic segmentation** task via pre-trained VL models

- **Goal:** To perform semantic segmentation only given a set of the target semantic category words
    - But **without** any task-specific images and annotations

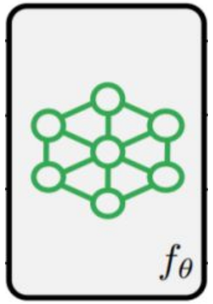


Contributions: Image-free Semantic Segmentation Task

We introduce a novel **image-free semantic segmentation** task via pre-trained VL models

- **Goal:** To perform semantic segmentation only given a set of the target semantic category words
  - But **without** any task-specific images and annotations
  - Target semantic categories: "grass" word "cat" word "dog" word

- **Solution**: We **replace** the **training dataset** with the artificially constructed **grid of category words**!



Pre-trained VL model

Image-free semantic segmentation

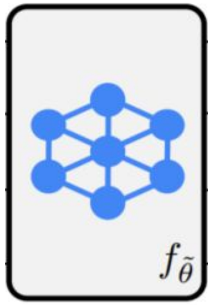
Requirements

cat	cat	cat	grass
grass	dog	grass	grass
dog	dog	dog	grass

Words Grid

cat	cat	cat	grass
grass	dog	grass	grass
dog	dog	dog	grass

Self-supervised Labels

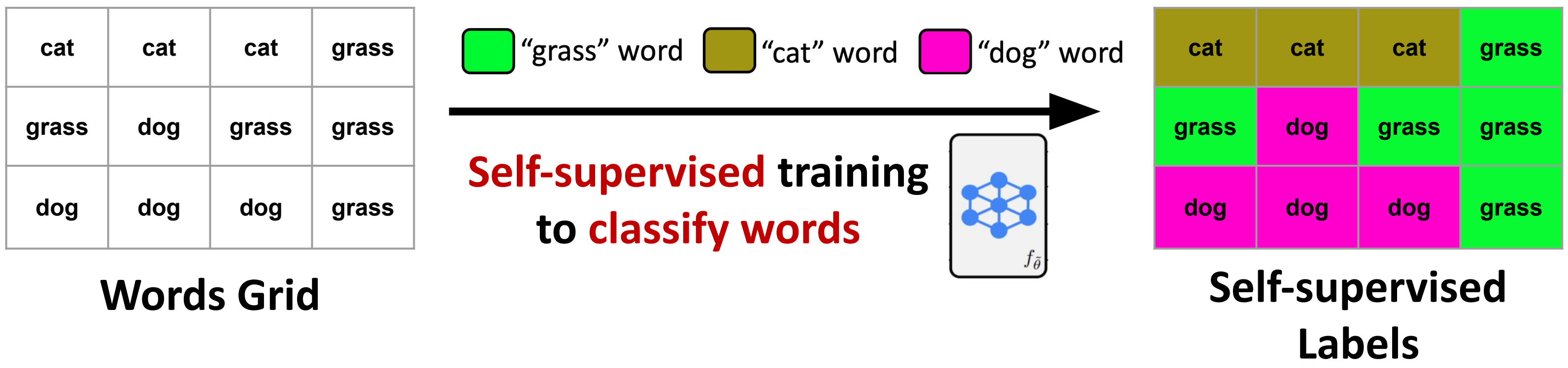


VL segmentation model

Contributions: Image-free Semantic Segmentation Task

Image-Free semantic **Seg**mentation (**IFSeg**) via Vision-Language (VL) models

- We propose a novel **self-supervision** method enabling **zero-shot semantic segmentation**
  - Learning to classify **category words** can adapt **pre-trained VL model** for image segmentation!
    no images, no dense annotations are required at all!

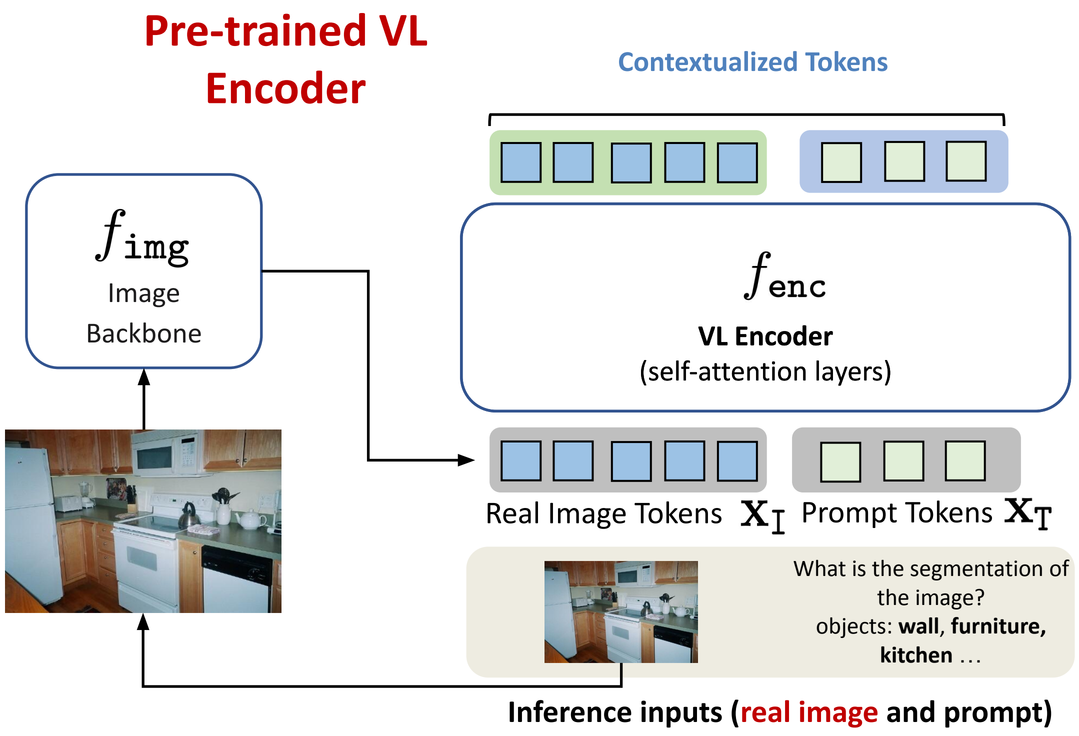


Zero-shot inference on real images

Method: image-free training

Key Idea: category word tokens can serve as image tokens on their **embedding space**

- **Cross-modal embedding space:** semantically similar {visual, word} tokens are closely located (*i.e.*, contextualized)

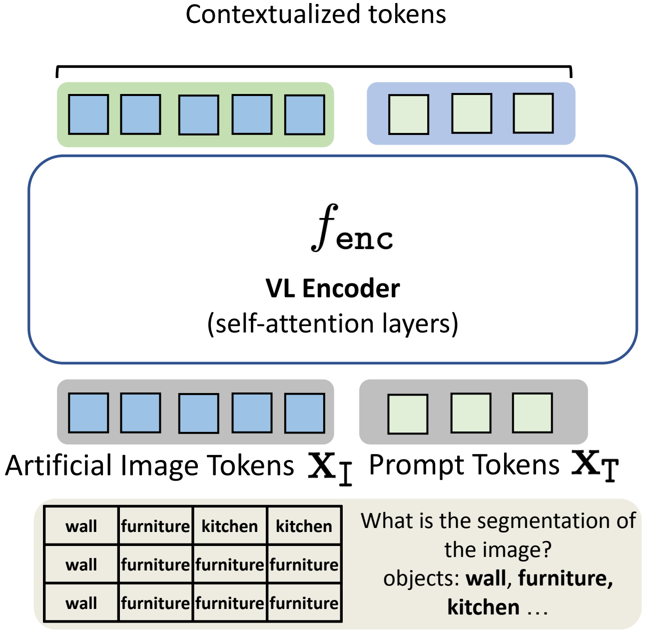


Method: image-free training

Key Idea: category word tokens can serve as image tokens on their **embedding space**

- **Cross-modal embedding space:** semantically similar {visual, word} tokens are closely located (*i.e.*, contextualized)

Training Time



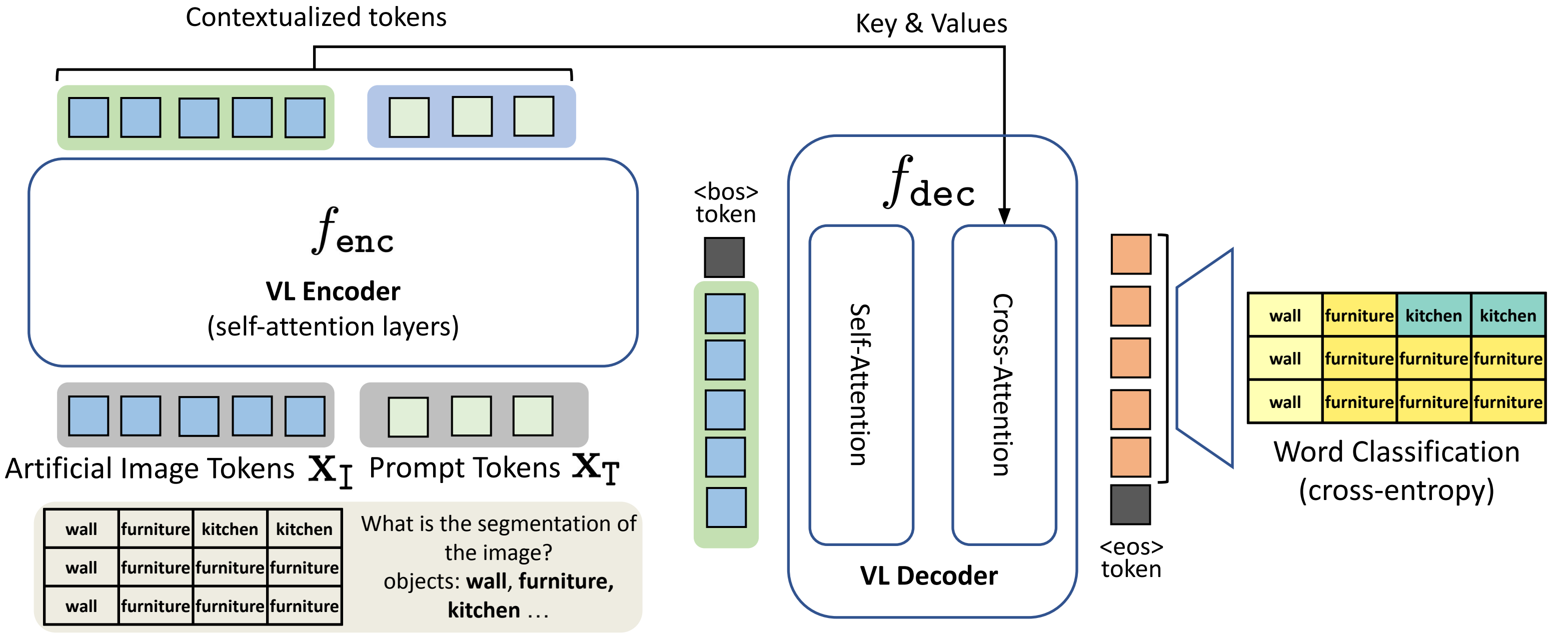
Training inputs (artificial image and prompt)

Method: image-free training

Key Idea: category word tokens can serve as image tokens on their **embedding space**

- **Cross-modal embedding space:** semantically similar {visual, word} tokens are closely located (*i.e.*, contextualized)
- **Requirements:** Then, we need a **VL decoder** to densely classify the tokens!
  - Therefore, we introduce a pre-trained **VL encoder-decoder** architecture and adapt it for the semantic segmentation

Training Time



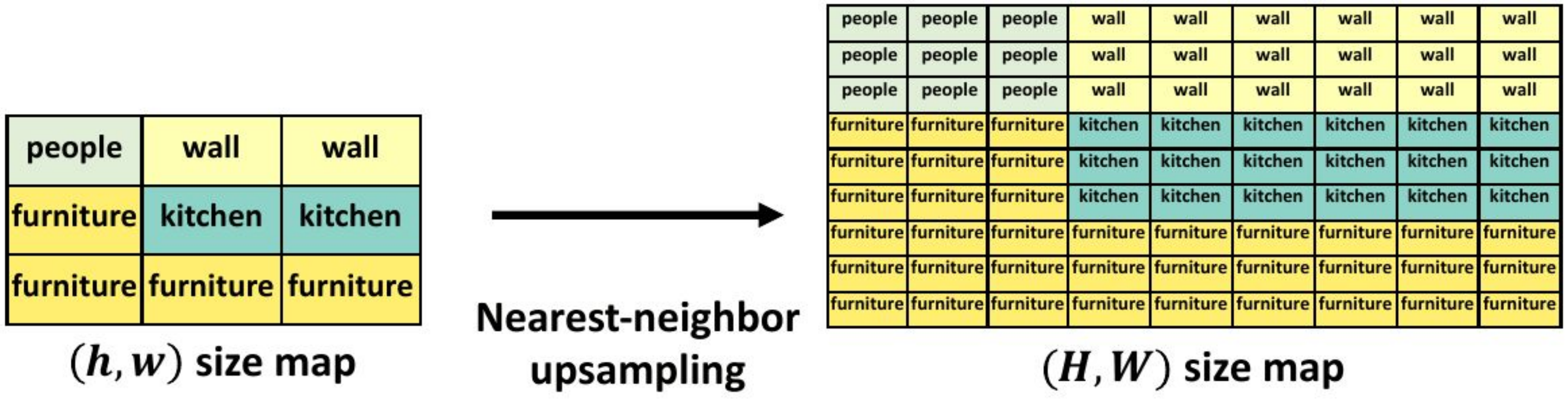
Training inputs (artificial image and prompt)

Method: artificial image (*i.e.*, word grid)

Artificial Image: 2D map of random semantic categories, constructed with the 2-step process

- In each training iteration, sample a grid of random size $(h, w) \sim \{1, 2, \ldots, s\}$
  - We use hyperparameter *s* = 32
- Then, upsample the variable-sized (h, w) to the fixed size of (H, W) using the nearest-neighbor algorithm
  - We use *H* = *W* = 32

- Our method ensures the **diversity of the inputs** while **regularizing the shapes**
  - **Real objects tend to be a cluster of various sizes** rather than being scattered
  - **Artificial image** empirically provides efficacy **comparable** to the **real segmentation masks**



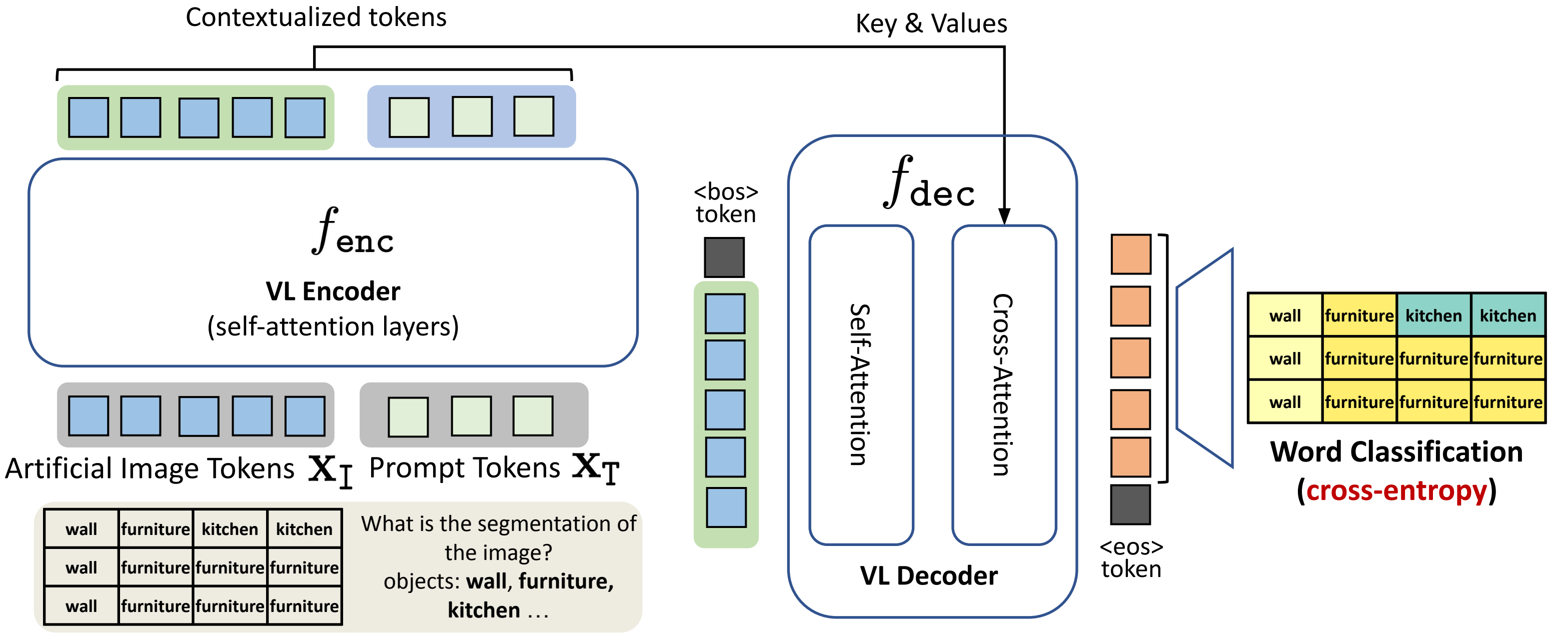
Method: image-free training

Key Idea: category word tokens can serve as image tokens on their **embedding space**

- **Cross-modal embedding space:** semantically similar {visual, word} tokens are closely located (*i.e.*, contextualized)
- **Requirements:** Then, we need a **VL decoder** to densely classify the tokens!
  - Therefore, we introduce the pre-trained **VL encoder-decoder** architecture and adapt it for the semantic segmentation

The encoder-decoder networks are **trained end-to-end using the cross-entropy loss** of the word classification!

Training Time

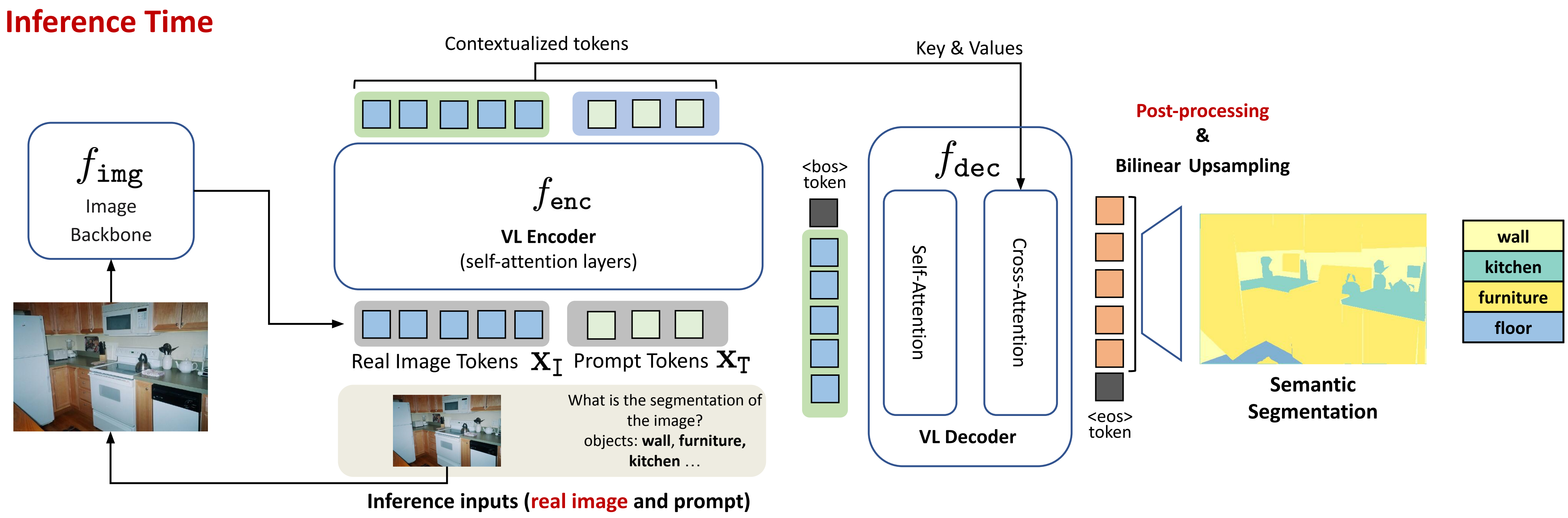


Training inputs (artificial image and prompt)

Method: zero-shot semantic segmentation

After tuning with the artificial image, the model is able to segment the target semantic categories

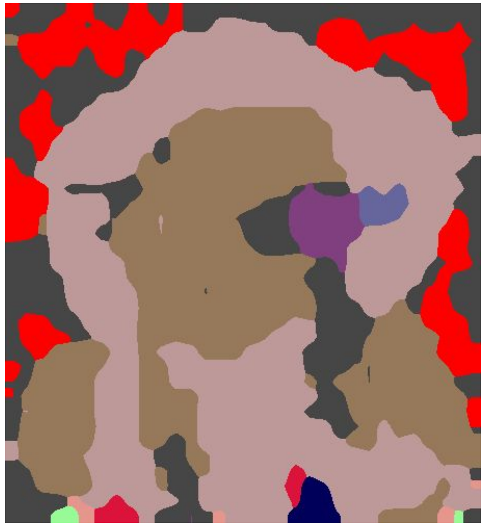
- The VL pre-trained image backbone (*e.g.*, ResNet) embeds the real image tokens
  - **Important**: image backbone remains frozen during IFSeg's training time!
- For realistic object shapes, we inject the shape extracted from the image features as **post-processing**
- Finally, the **bilinear upsampling** resizes the output to a desired shape



Method: post-processing

- **Challenge of image-free segmentation:** Discrepancy of input modality between training and evaluation
  - Due to the absence of training images
- We design visual feature-based post-processing technique for better semantic segmentation
  1. We first search K-nearest neighbors with the cosine distance for each image feature
     - On the frozen output embeddings of image backbone network
  2. We average corresponding neighbors on the predictive distributions of the VL decoder

Image



Segmentation Result (before)

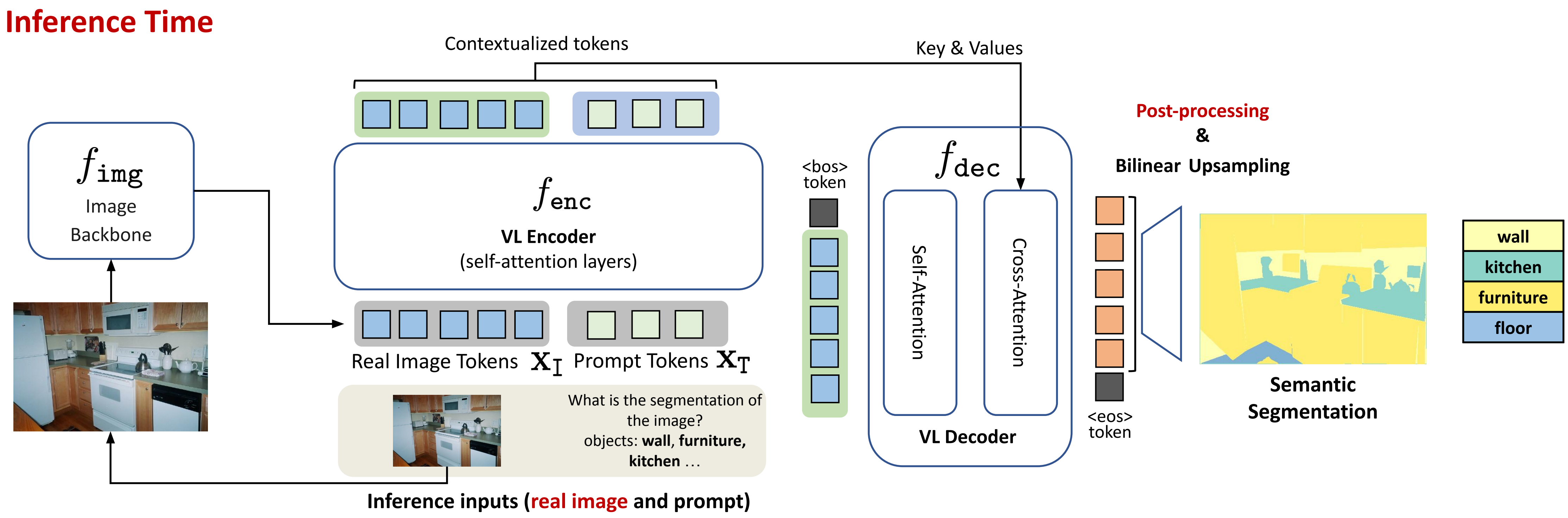
Post-processing →

Segmentation Result (after)

Method: zero-shot semantic segmentation

After tuning with the artificial image, the model is able to segment the target semantic categories

- The VL pre-trained image backbone (*e.g.*, ResNet) embeds the real image tokens
  - **Important**: image backbone remains frozen during IFSeg's training time!
- For realistic object shapes, we inject the shape extracted from the image features as **post-processing**
- Finally, the **bilinear upsampling** resizes the output to a desired shape



Quantitative Results

Comparison with VL-baselines under two different zero-shot scenarios

- Benchmark in COCO-stuff semantic segmentation, using the **mean intersection over union (mIoU)** metric
- **Annotation-free and Image-free** scenario (**left**)
  - **+30.8** mIoU than the image-free, and **+6.9** mIoU than image-aware baselines (MaskCLIP / MaskCLIP+)
- **Weakly-supervised**: 156 “seen” classes for training / 15 “unseen” classes for zero-shot inference (**right**)
  - **+2.1** mIoU on unseen classes, “mIoU(U)” than the strongest baseline (MaskCLIP+)

IFSeg archives a new mIoU record on both the **image-free** and the **Weakly-supervised** scenarios!

Method	Backbone	Image Dataset	mIoU
MaskCLIP+ [53]	ResNet-101	COCO (118k)	48.7
CLIP [32, 53]	ResNet-101	✗	12.3
OFA [42]	ResNet-101	✗	6.8
MaskCLIP [53]	ResNet-101	✗	24.8
IFSeg (ours)	ResNet-101	✗	**55.6**

Image-free
Semantic Segmentation

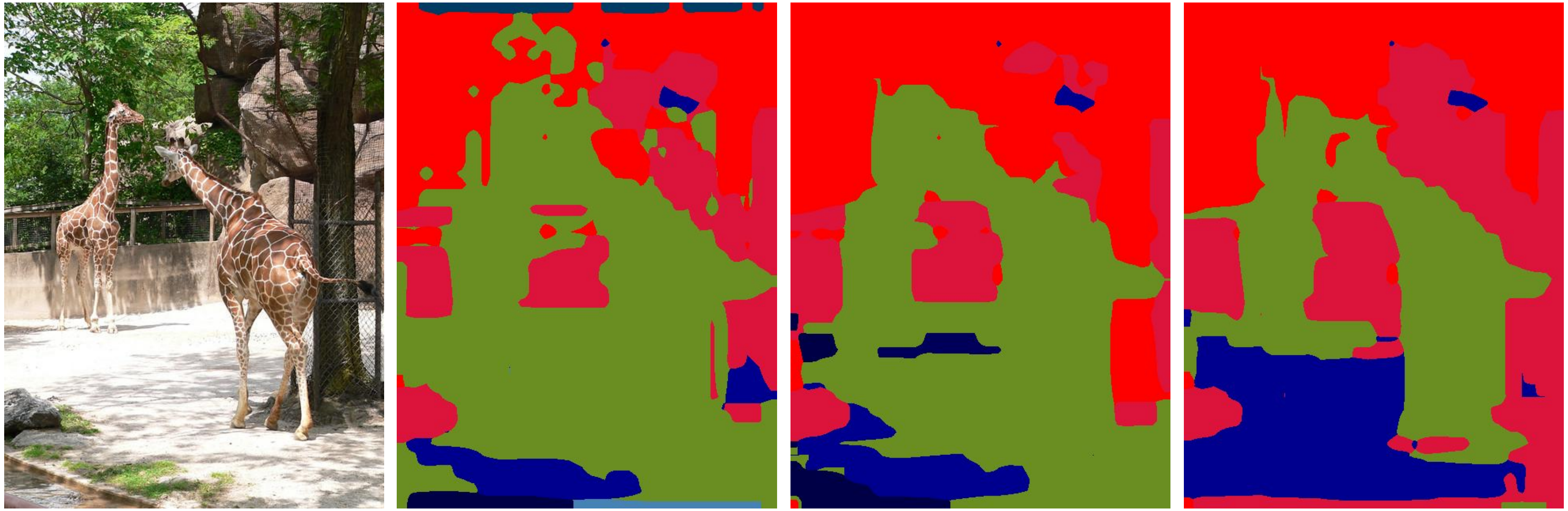
Method	Text Backbone	Image Backbone	Image Dataset	Segmentation Label	mIoU(U)	mIoU(S)	hIoU
ZS5 [4]	word2vec [29]	ResNet-101	COCO (118k)	✓(156 seen)	10.6	34.9	16.2
CaGNet [16]	word2vec [29], fasttext [21]	ResNet-101	COCO (118k)	✓(156 seen)	13.4	35.3	32.6
SIGN [7]	word2vec [29], fasttext [21]	ResNet-101	COCO (118k)	✓(156 seen)	15.2	36.4	21.4
SPNet [45]	word2vec [29], fasttext [21]	ResNet-101	COCO (118k)	✓(156 seen)	26.9	34.6	30.3
STRICT [31]	word2vec [29], fasttext [21]	ResNet-101	COCO (118k)	✓(156 seen)	30.3	35.3	32.6
ZSSeg [48]	ALIGN-BERT-Large [20]	ResNet-101	COCO (118k)	✓(156 seen)	43.6	39.6	41.5
MaskCLIP+ [53]	CLIP-ResNet [32]	ResNet-101	COCO (118k)	✓(156 seen)	54.7	38.2	45.0
IFSeg+ (ours)	OFA-Base [32]	ResNet-101	COCO (118k)	✓(156 seen)	**56.8**	**41.9**	**48.2**

Weakly-supervised
Semantic Segmentation

Ablation Study: notes on the artificial image

Random sampling in the artificial image construction

- Empirical performance in terms of mIoU: Deterministic Shape < **Ours** ≤ GT
  - Our **Random Shape + upsampling** approach is comparable to using GT mask as the artificial image!



Input Image Deterministic Shape Ours GT Mask

(H, W)	(h, w)	mIoU
32	Deterministic	47.7
8	$(h, w) \sim \{1, 2, ..., 8\}$	55.8
16	$(h, w) \sim \{1, 2, ..., 16\}$	57.8
32	$(h, w) \sim \{1, 2, ..., 32\}$	55.6
32	Ground Truth	**58.5**

Effect of random sampling in artificial image

Summary

We introduce **Image-free semantic segmentation (IFSeg)** via the pre-trained VL models

- **Key idea:** Semantic categories can serve as artificial image tokens in the cross-modal latent space
- We propose 2D map of random semantic categories as artificial image to train the model in an image-free manner
  - Without a burden of acquiring additional training images or even segmentation annotations

- IFSeg is an effective baseline for the image-free semantic segmentation task
  - We also highlight the broad applicability of this task to evaluate tending VL models

Poster: TUE-AM-282

GitHub: https://github.com/alinlab/ifseg

Paper: https://arxiv.org/abs/2302.14115

KAIST

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Google