Self-Attention-Driven Refinements for Training-Free Segmentation in Diffusion Transformers

Rento Yamaguchi, Keiji Yanai
The University of Electro-Communications, Tokyo, Japan

Introduction



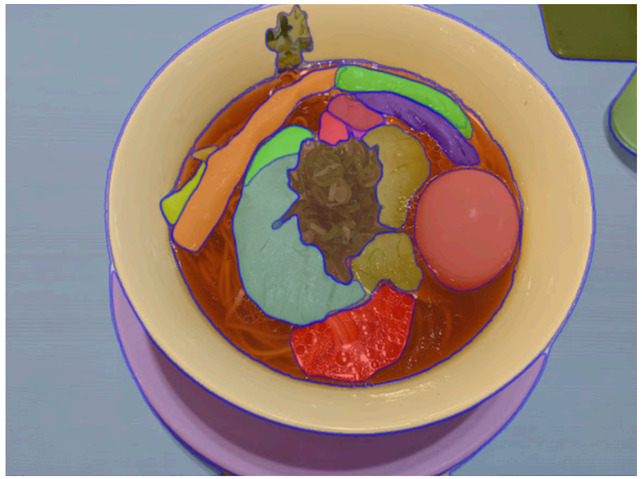
Research Background

Semantic Segmentation is a task that involves
predicting what is shown in each pixel of an image.

Training segmentation models requires a large amount of data.

➡ Collecting paired data and training the models can be expensive.





A large data
is needed

Research Purpose

Research Challenges

Training models requires
a large amount of training data

Research Purpose

Research Challenges

Training models requires
a large amount of training data

Using a large pre-trained model to create
a segmentation model
that does not require additional data or training

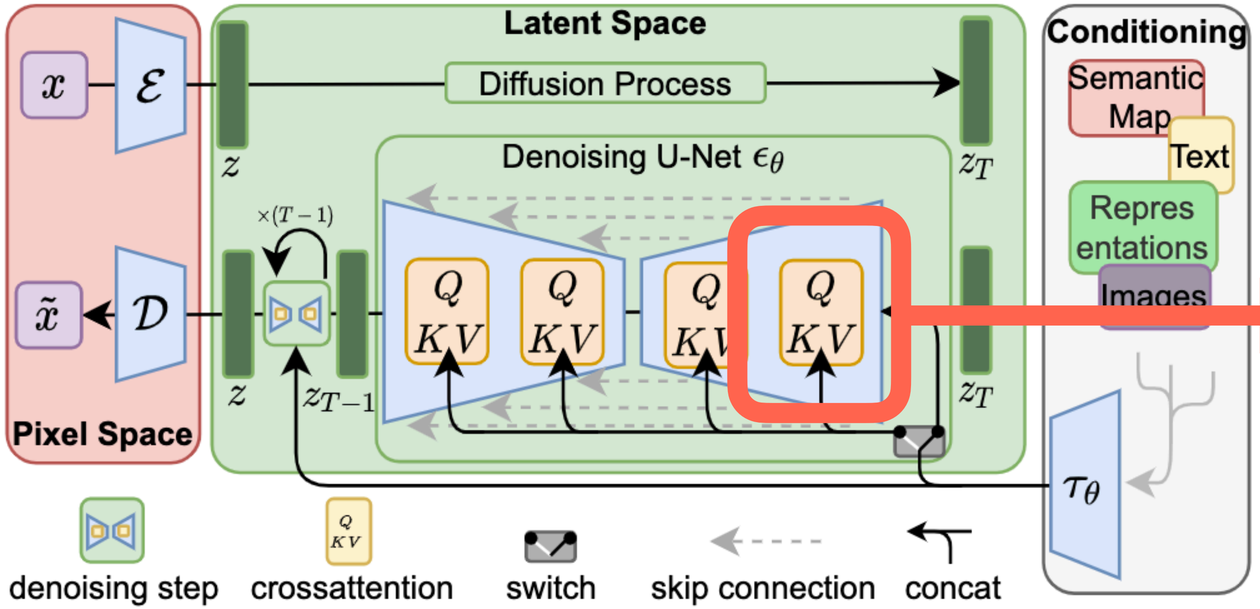
Related Works



Diffusion Models (DM)

[1] High-resolution image synthesis with latent diffusion models
Robin Rombach et al. CVPR, 2022

Models like Stable Diffusion can generate high-quality images and are trained on large datasets of text and images



Overview of the model used in Stable Diffusion [1]

To create images from text, a mechanism to handle the relationship is necessary

The U-Net's cross-attention mechanism is used for conditioning on the text

Attention Maps

[2] Prompt-to-prompt image editing with cross attention control
Amir Hertz et al. arXiv:2208.01626, 2022

Stable Diffusion use the cross-attention during generation

The cross-attention map captures
the approximate location of objects

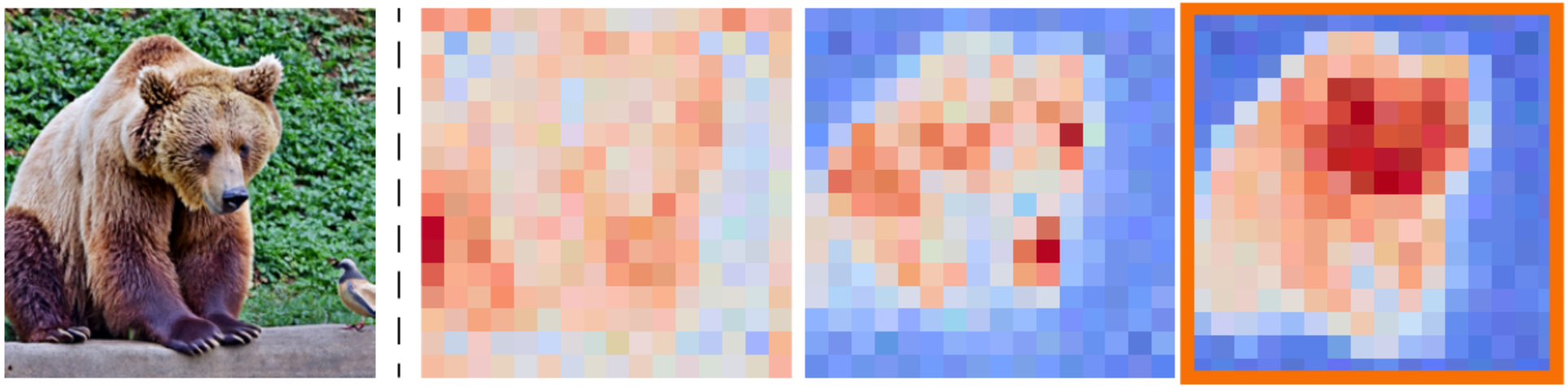
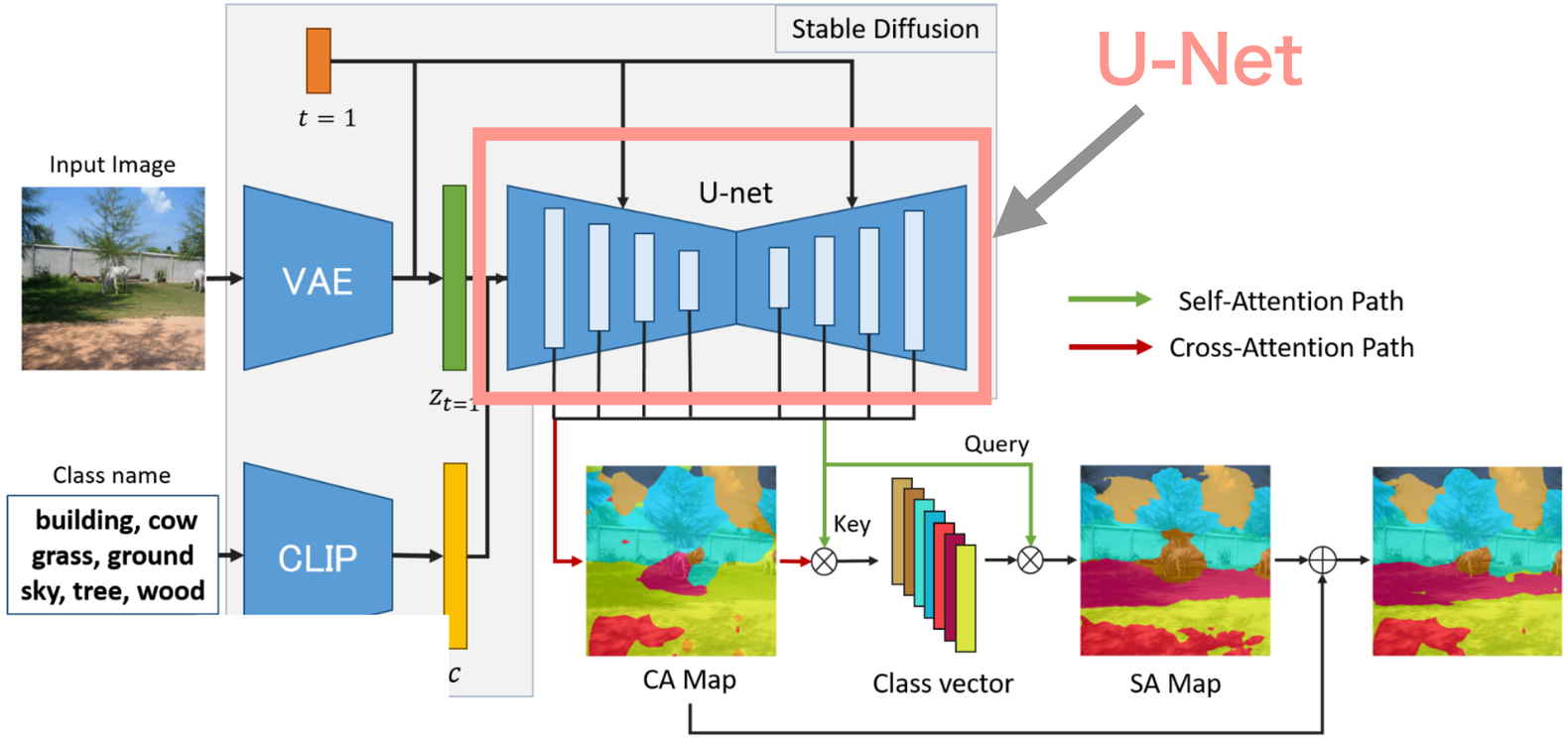


Image generation with 'a furry bear' and visualization of the cross-attention map corresponding to each word [2]

Segmentation Using Diffusion

[3] Training-Free Region Prediction with Stable Diffusion
Yuma Honbu et al. MMM 2023

Methods like StableSeg [3] utilize attention maps to achieve zero-shot segmentation

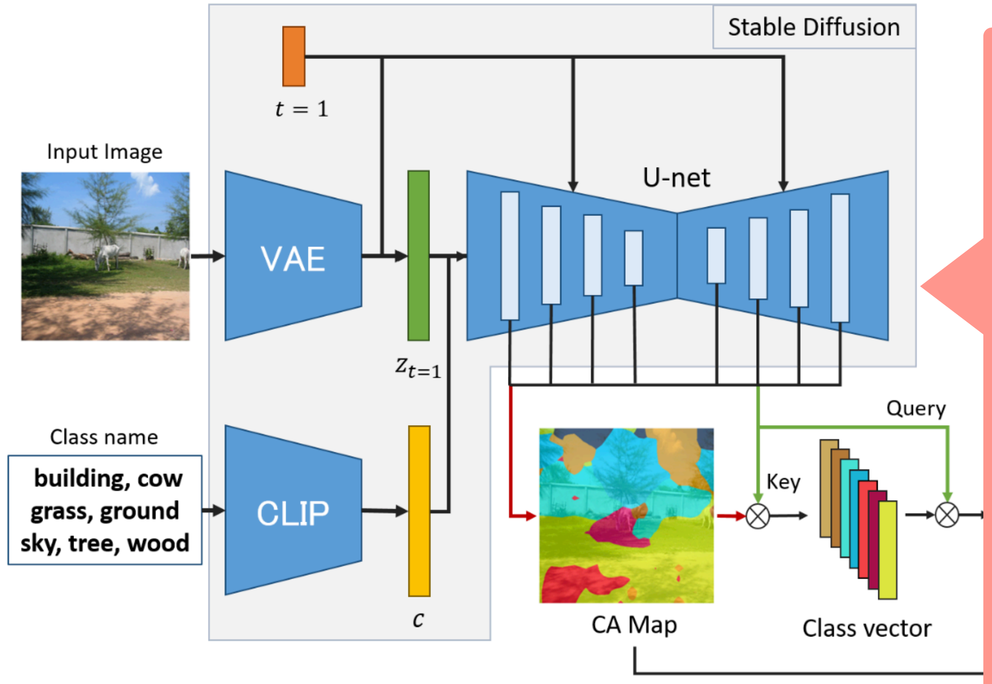


Segmentation Using Diffusion

[3] Training-Free Region Prediction
with Stable Diffusion
Yuma Honbu et al. MMM 2023

Existing methods that use diffusion models are all U-Net based.

➡ This approach cannot be directly applied to diffusion models with different structures.

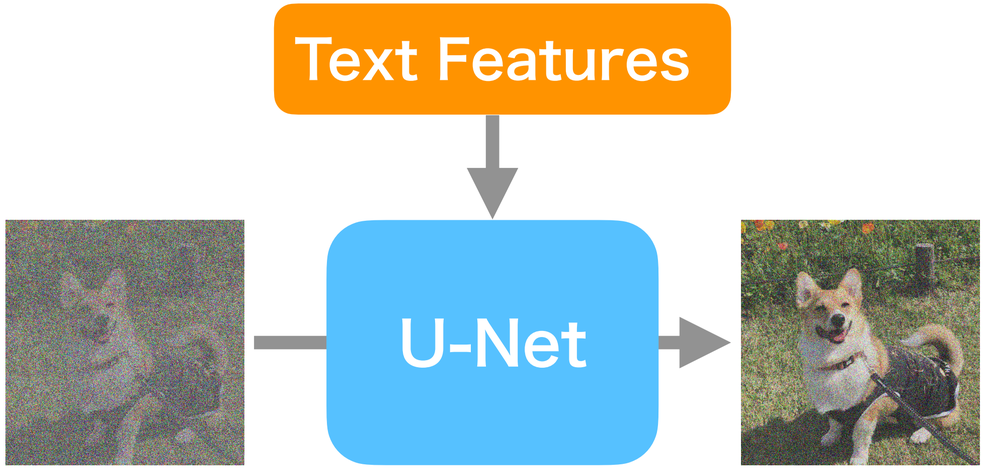


Apply the cross-attention maps obtained from the U-Net model, which relate text to image, to segmentation tasks

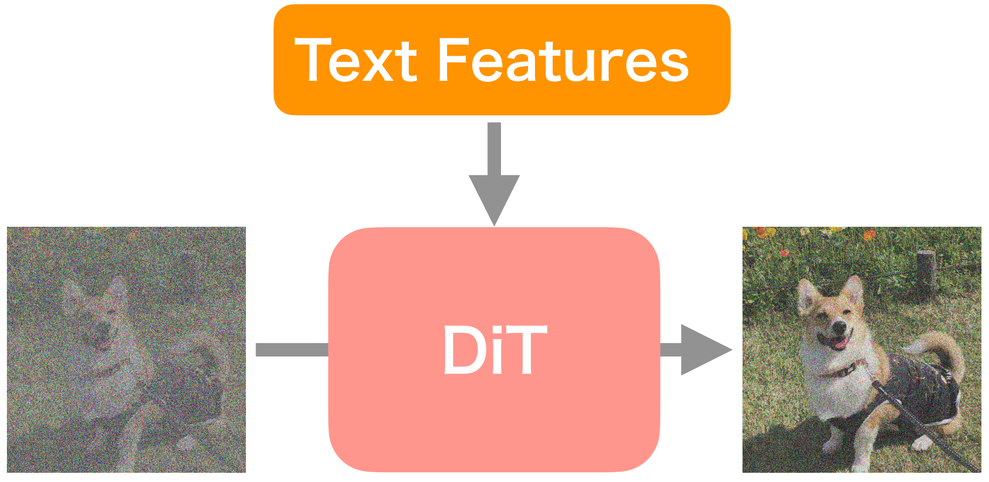
Advancements in Diffusion Models

Recently, Diffusion Transformer (DiT)-based diffusion models have enabled even higher-quality image generation

U-Net based model



DiT based model



Advancements in Diffusion Models

Recently, Diffusion Transformer (DiT)-based diffusion models have enabled even higher-quality image generation

Input prompt:
“A cat holding a sign that says Hello World”

U-Net based model



Stable Diffusion v1.4

DiT based model



Stable Diffusion 3



FLUX.1-dev

Advancements in Diffusion Models

Q: Could DiT-based models, which can generate higher-quality images than traditional models, enable more accurate segmentation?

Input prompt:
“A cat holding a sign that says Hello World”

U-Net based model

DiT based model



Stable Diffusion v1.4



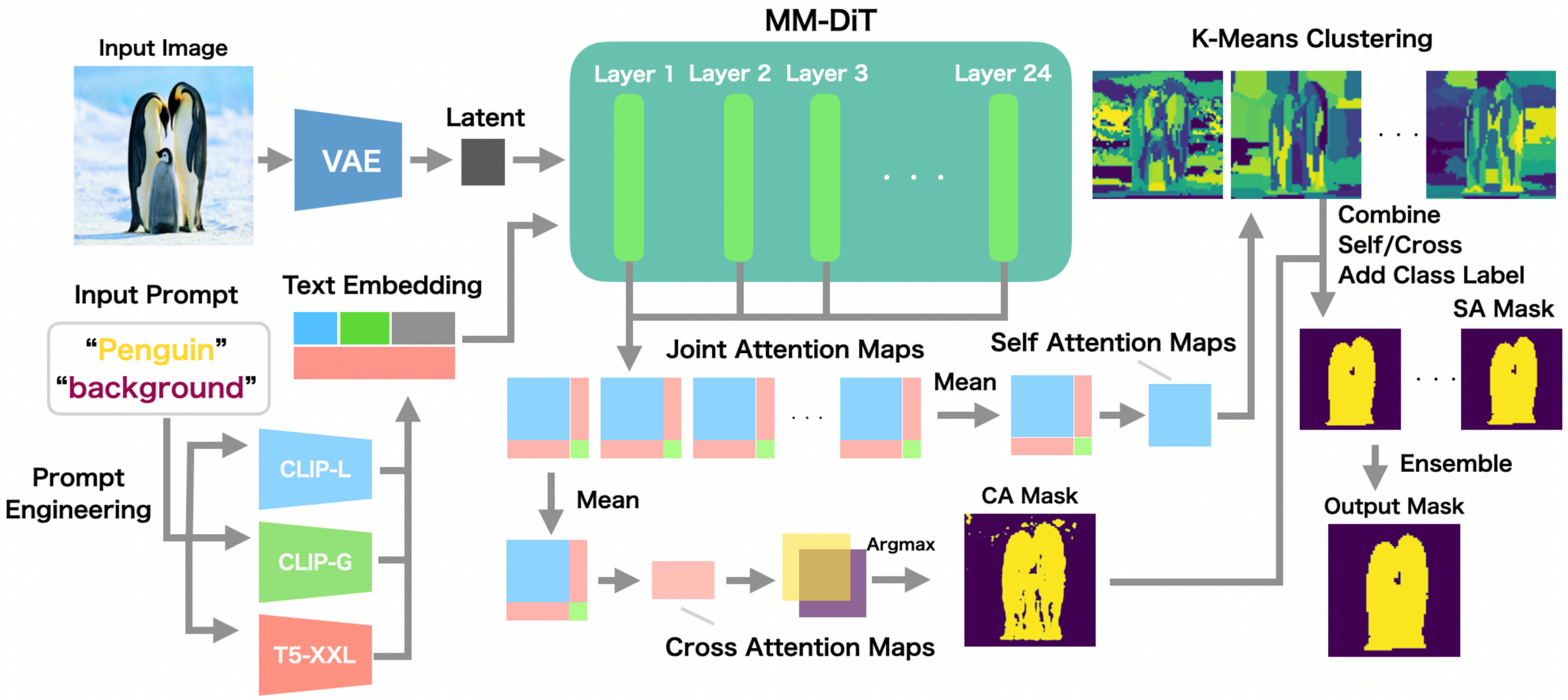
Stable Diffusion 3



FLUX.1-dev

Methodology

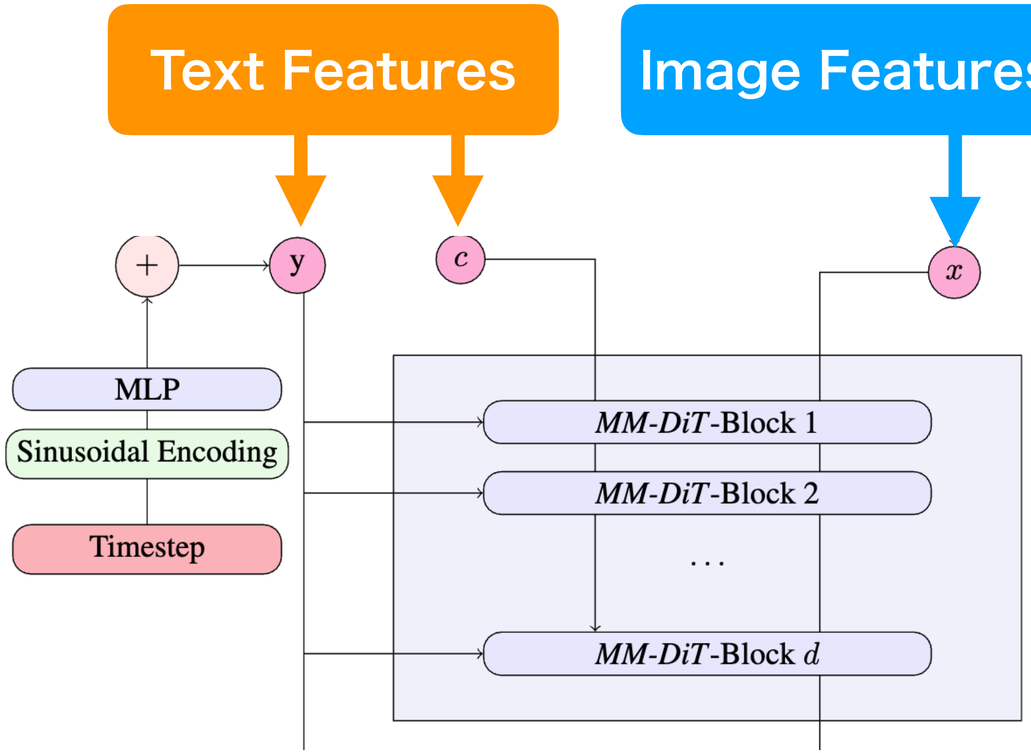
Overview Diagram of the Proposed Method



Stable Diffusion 3

[3] Scaling Rectified Flow Transformers for High-Resolution Image Synthesis
Patrick Esser et al., ICML, 2024

Perform zero-shot segmentation using Stable Diffusion 3 [3], which incorporates the Multi-modal Diffusion Transformer (MM-DiT)



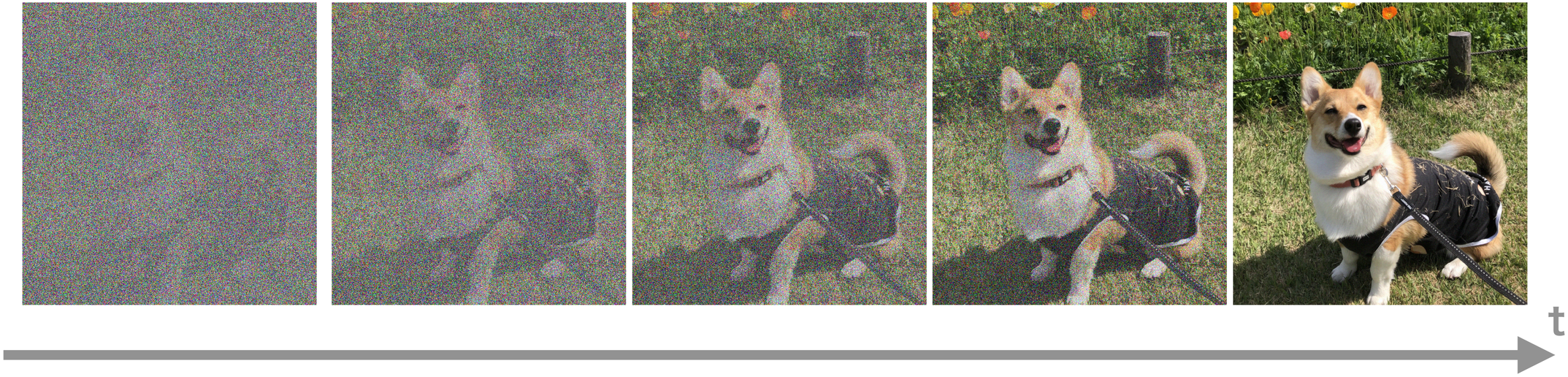
Different Structure from Traditional SD Using U-Net

Existing Methods Cannot Be Applied Directly

Proposed Method Setup

Diffusion models generate images by iteratively removing noise

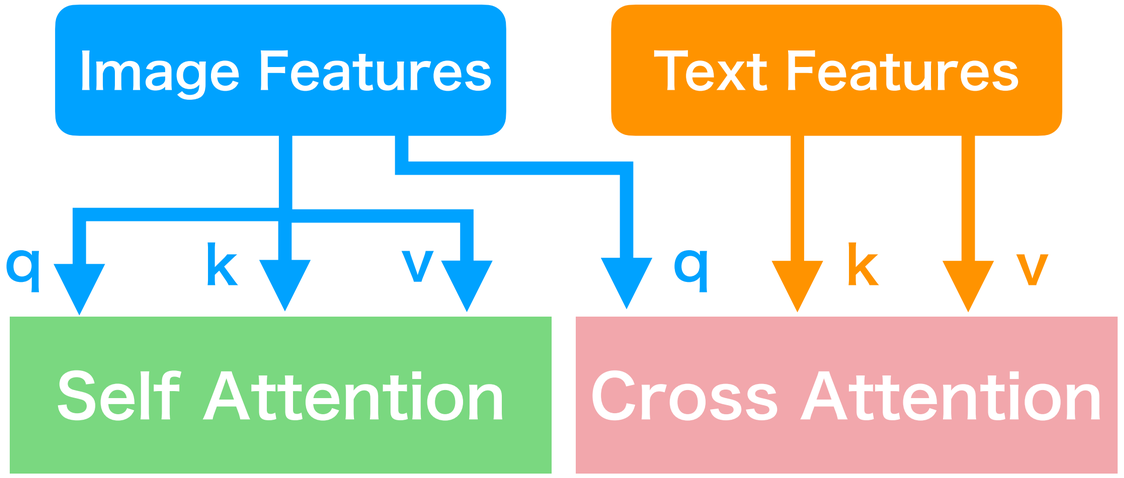
➡ Following existing research, assume a specific single step in MM-DiT to obtain the attention maps for text and images



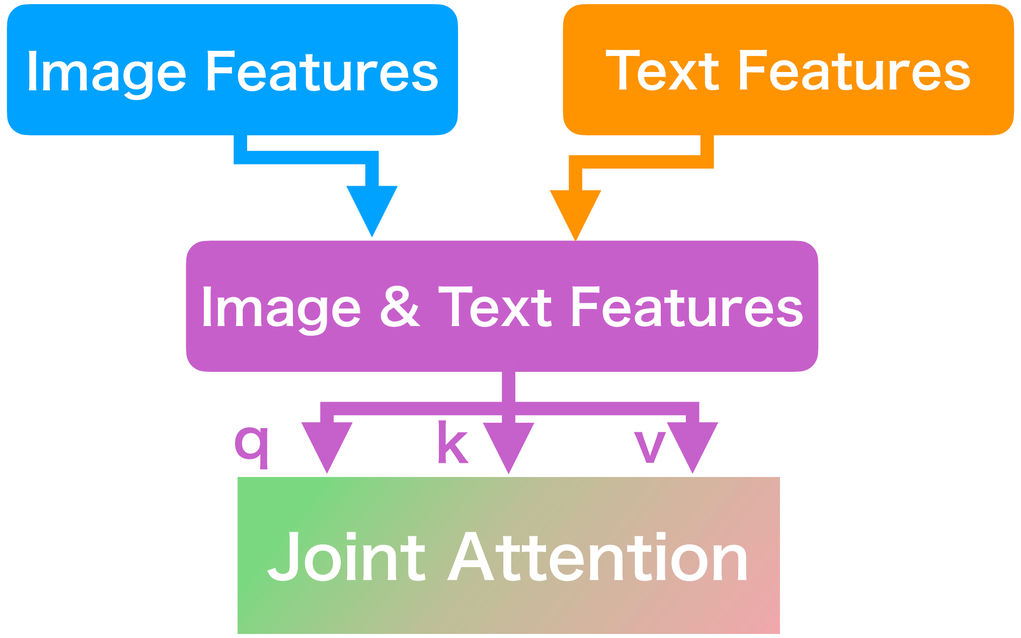
Attention Mechanisms in U-Net and MM-DiT

Obtaining Cross-Attention Maps from Image and Text Features

The attention block in MM-DiT differs structurally from traditional Stable Diffusion (SD)



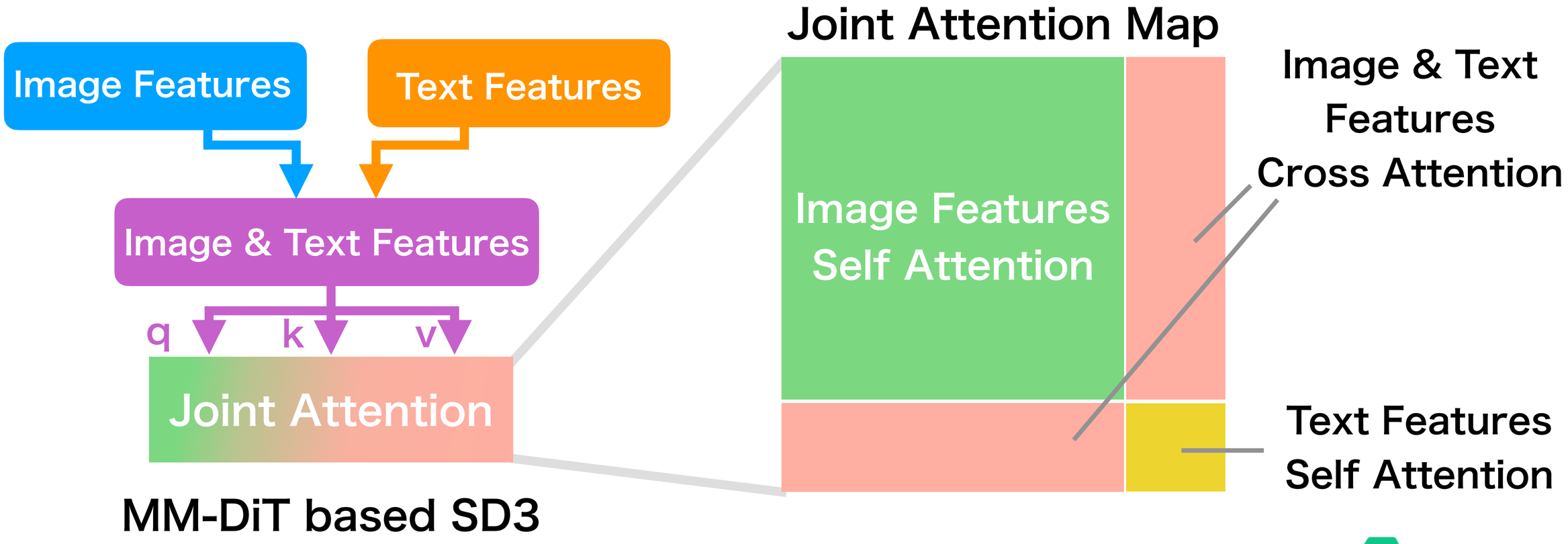
U-Net based SD



MM-DiT based SD3

Extraction of Attention Map from MM-DiT

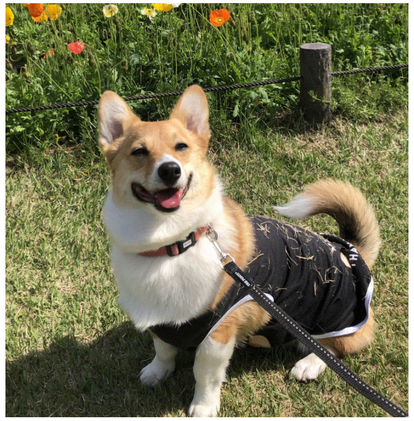
Extract Cross/Self Attention Maps with image and text information from Joint Attention in MM-DiT



Mask Generation from Cross Attention

Extract the Cross Attention for "dog" and "background" tokens and determine the class with the strongest response for each pixel

Input Image



Input Text

“dog”

“background”

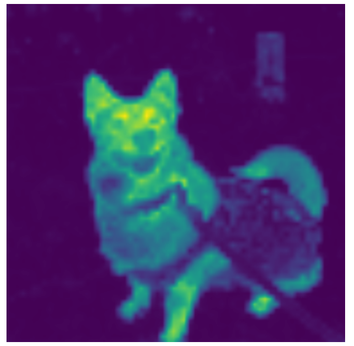
Cross Attention Map

Text Features

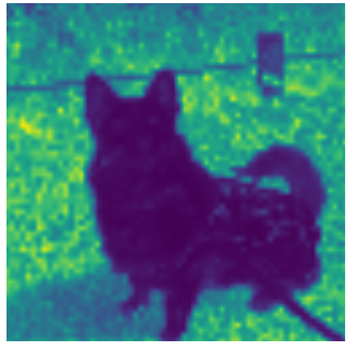
x Image Features

Mean

“dog”



“background”



argmax

CA Mask



with some noise

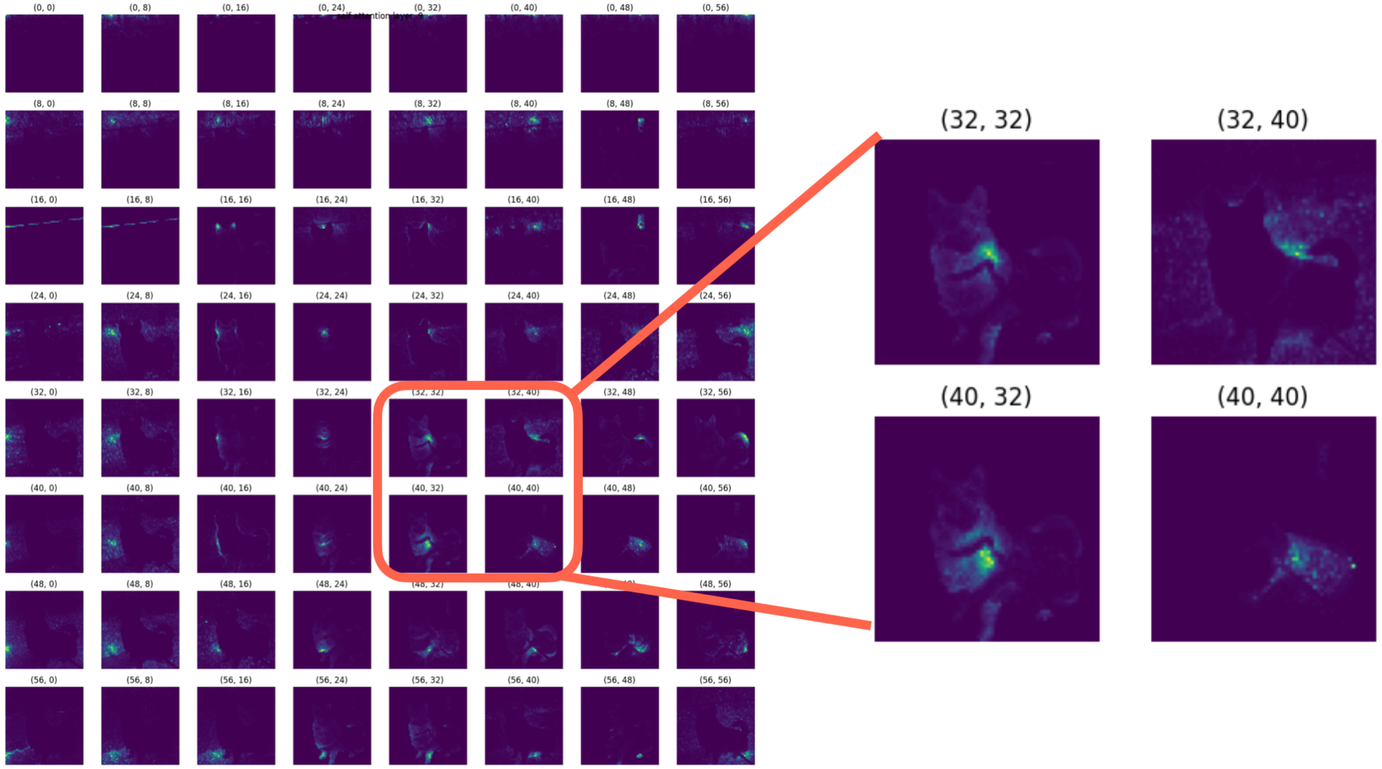
Self Attention Map

Self Attention Maps show responses
from pixels that are similar to each other

e.g.) dog's pixel ↔️ other dog's pixels

Input Image





Self Attention Map

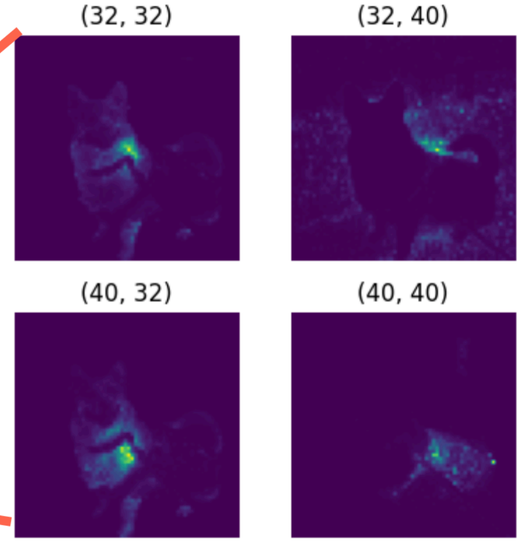
Self Attention Maps show responses
from pixels that are similar to each other

e.g.) dog's pixel ↔️ other dog's pixels

Input Image

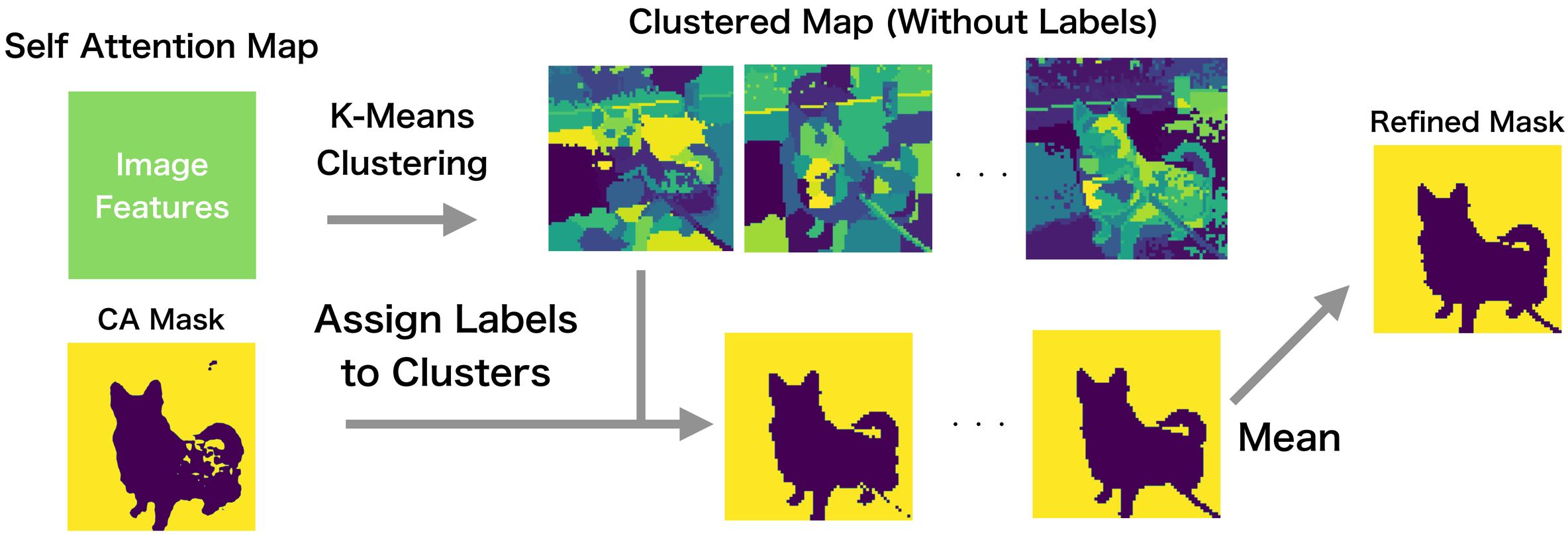


This map seems
capable of
capturing the
outlines of objects!



Refinement of Mask Using Self Attention

Extract Self Attention of image features and use K-Means to cluster similar pixels together



Assigning Class Labels

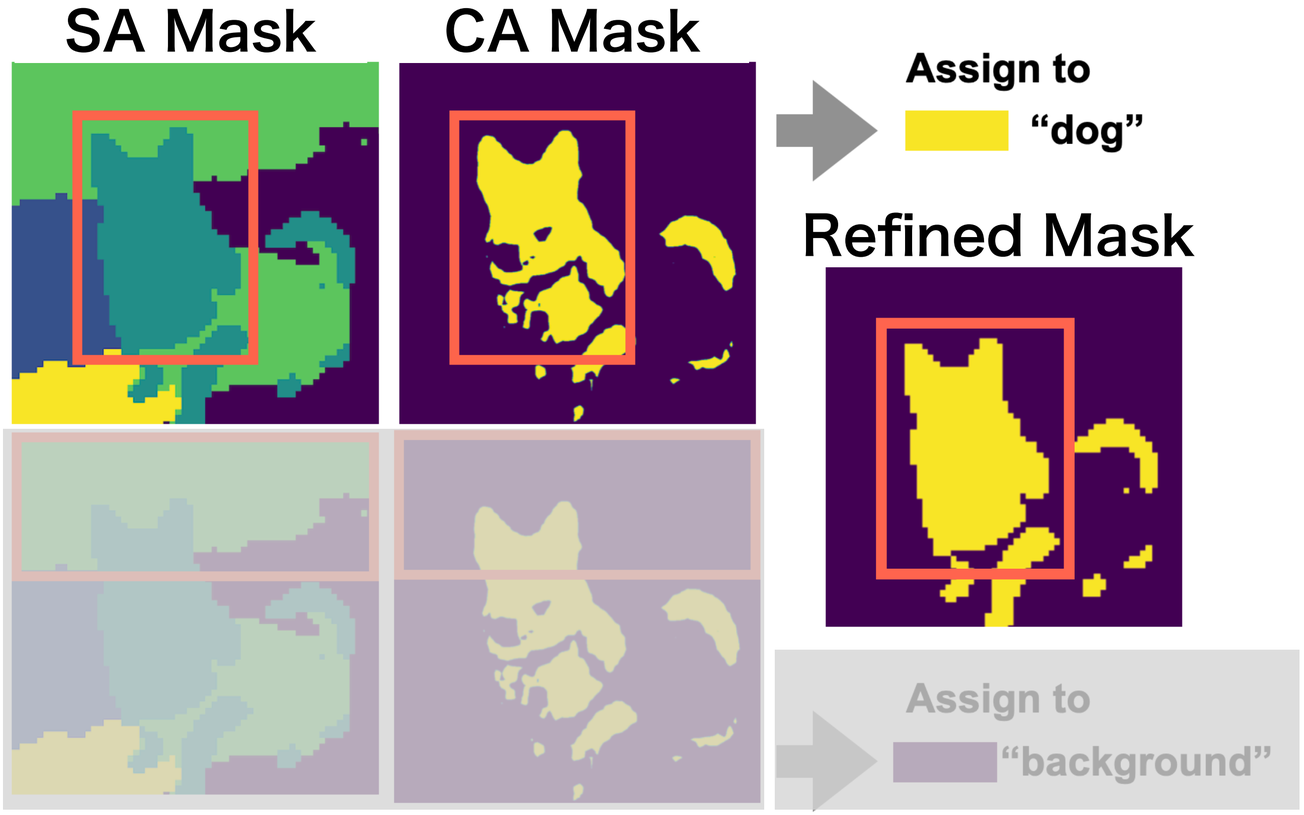
Compare the area occupied by each pixel in the clusters of the SA Map obtained from K-Means with the CA Mask and assign labels accordingly

Input Image



“dog”

“background”



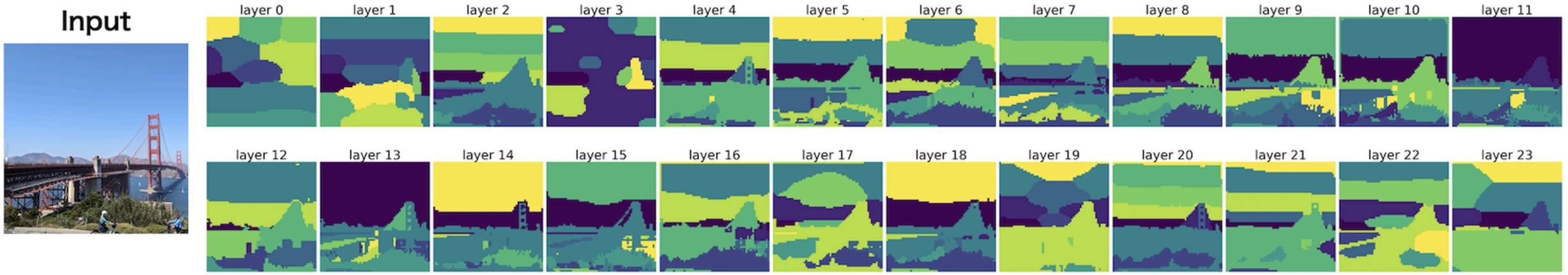
Experiments



Layer-wise Self Attention Analysis

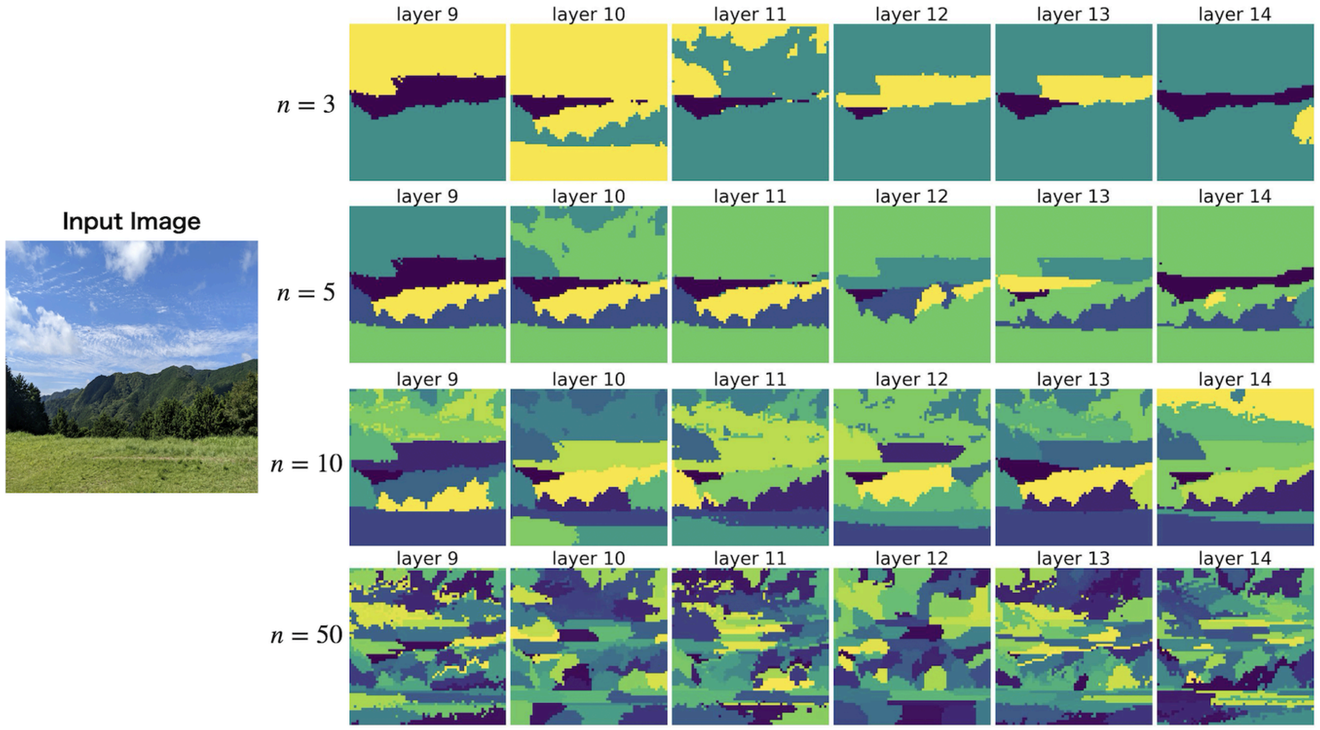
Extracted self-attention maps from MM-DiT and applied K-Means clustering to each, using a cluster number of n = 10

- Early and late layers show clustering not aligned with image object boundaries
- In intermediate layers, clusters may encompass different objects



Comparison on Number of Clusters

Increasing N of clusters results in finer segmentation of object boundaries for more accurate shape extraction
However, too many clusters can lead to over-clustering, where noise regions may receive inappropriate class labels



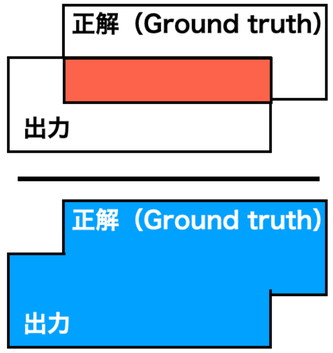
N_clusters	mIoU
10	0.448
20	0.461
30	0.462
50	0.462
100	0.461
200	0.459
w/o SA	0.452

Evaluation Using Segmentation Datasets

Evaluation on PASCAL VOC and Cityscapes Test Data
Achieved performance comparable to existing research, but slightly lower (mIoU ↑)

MaskCLIP:
  Existing Research Based on CLIP
StableSeg:
  Existing Research Using U-Net Based SD
Proposed Method:
  Approach Using DiT-Based SD

$$\text{IoU} = \frac{\text{出力と正解の積}}{\text{出力と正解の和}} =$$



	PASCAL VOC	Cityscapes
MaskCLIP	44.7	21.6
StableSeg	50.3	15.0
Ours	46.2	13.9

Comparison of Refinement Using Self Attention

Evaluate the increase in metrics by comparing the effect of mask refinement with existing studies

StableSeg:
Existing Research Using U-Net Based SD

Proposed Method:
Approach Using DiT-Based SD

Input



w/o SA



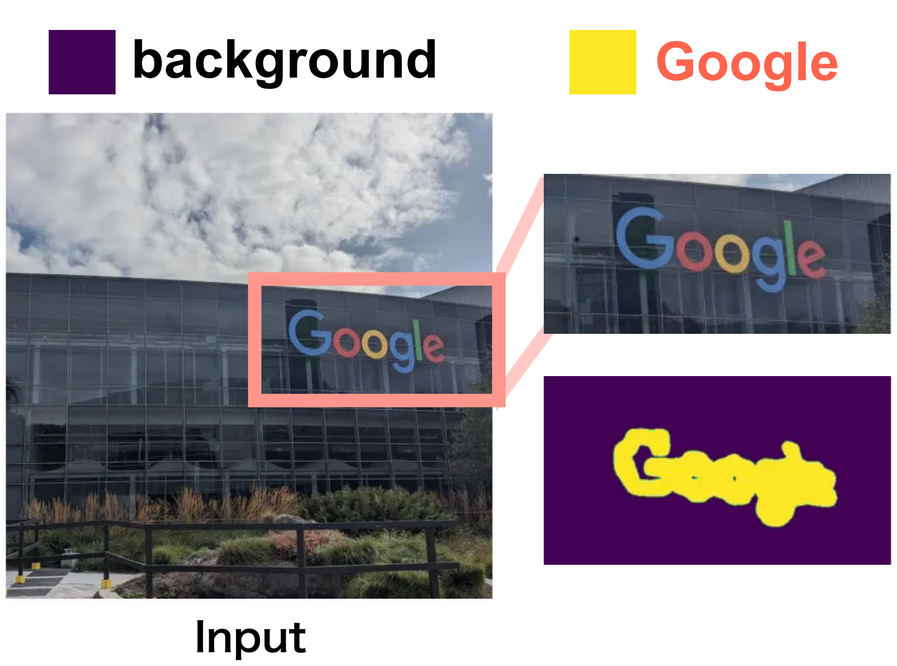
w/ SA

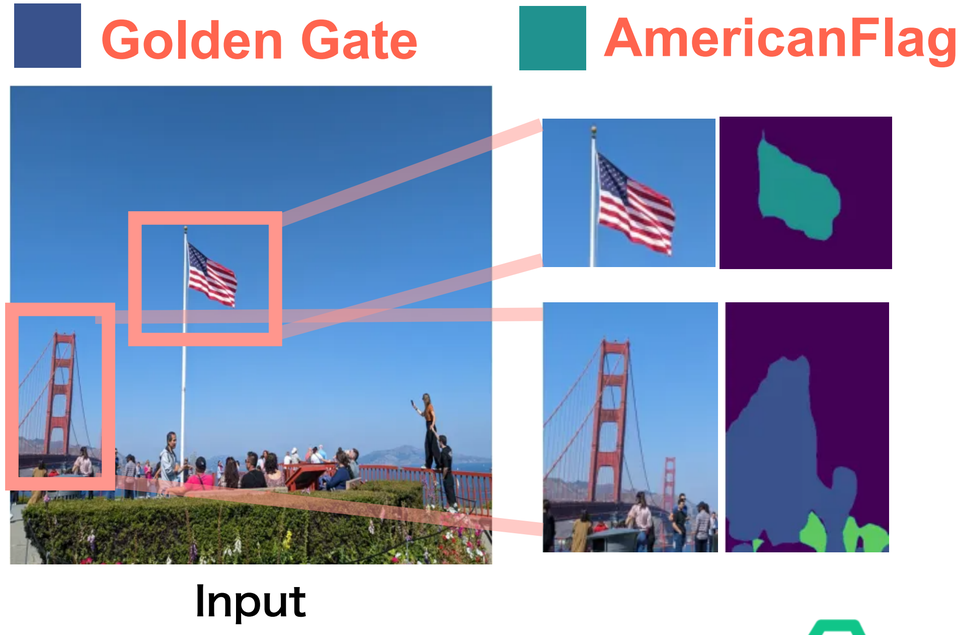


	PASCAL VOC	Cityscapes
StableSeg (w/o SA)	47.2	12.7
	↓ +3.1	↓ +2.3
StableSeg (w/ SA)	50.3	15.0
Ours (w/o SA)	45.2	13.7
	↓ +1.0	↓ +0.2
Ours (w/ SA)	46.2	13.9

Open-Vocabulary Segmentation

Stable Diffusion 3 is pre-trained on a large dataset from the web, enabling it to handle class labels for proper nouns and other specific terms using prior knowledge





Conclusion



Conclusion




Accomplishments

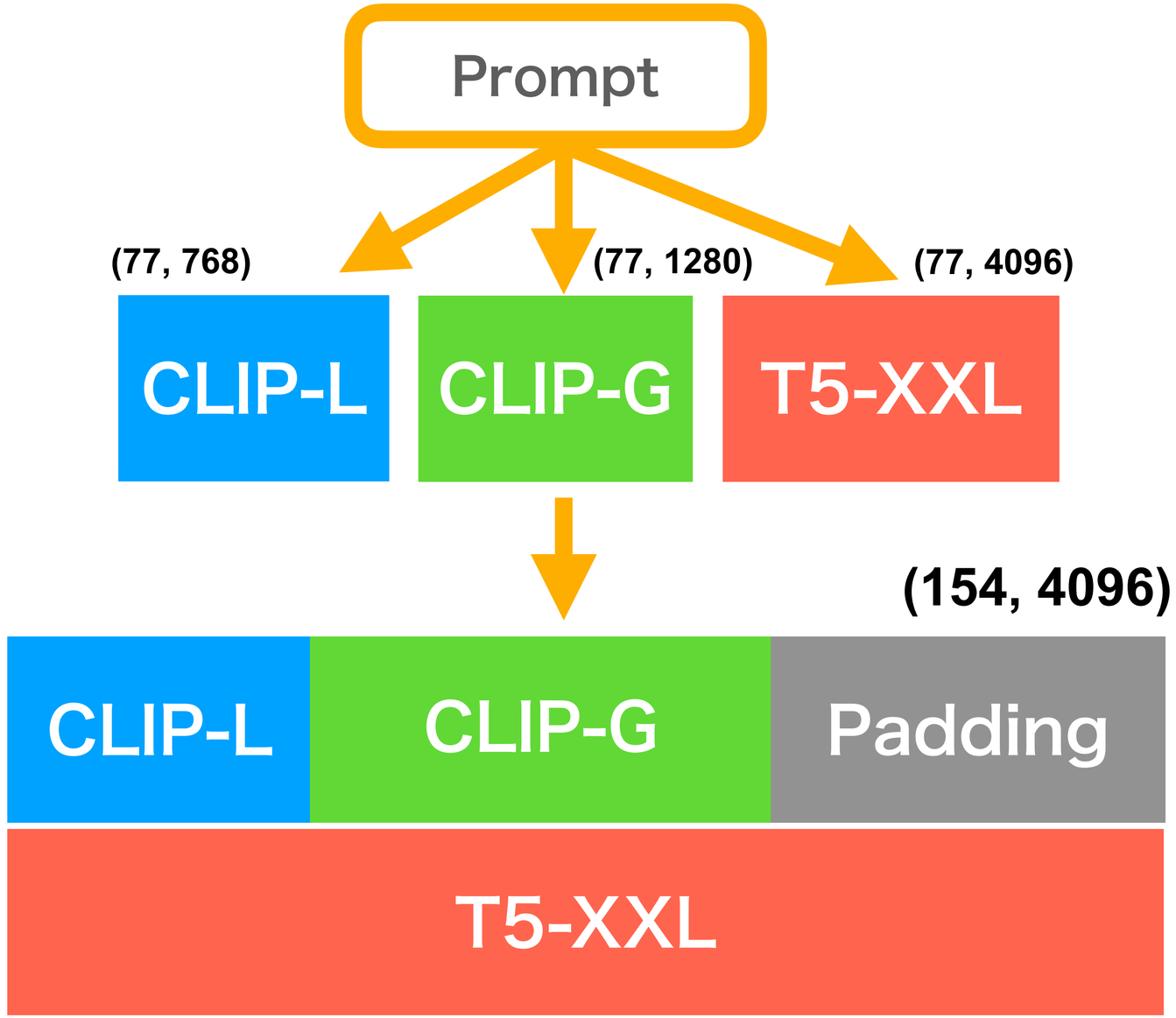
- Proposed a segmentation method without training using models with a DiT-based architecture (specially Self Attention Refinements)
- Currently, results are slightly inferior to existing methods

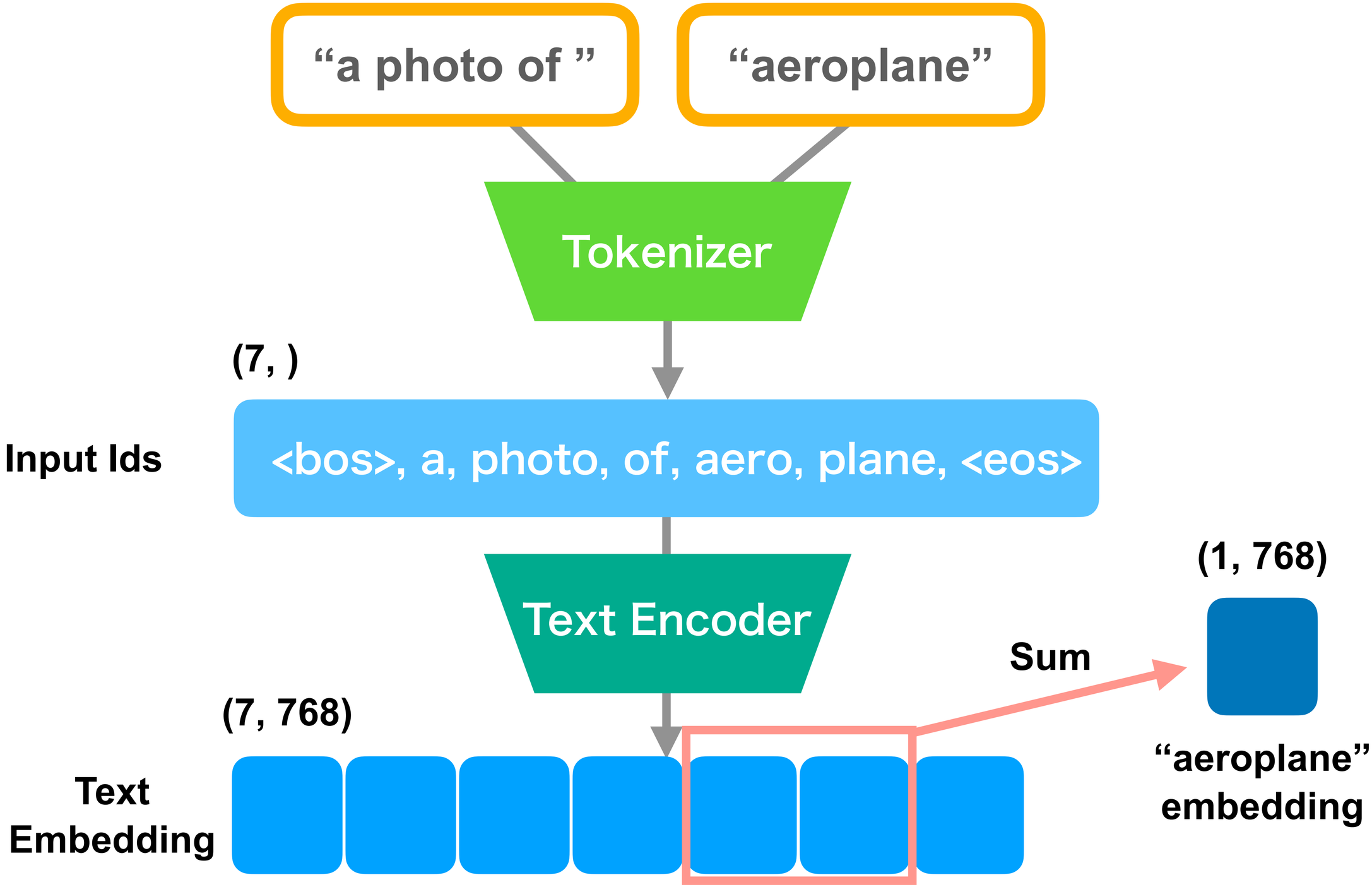
Future Challenges

- Compare the refinement methods using self-attention (SA) in existing approaches
- Investigate the disparity between high generation quality and lower segmentation accuracy.

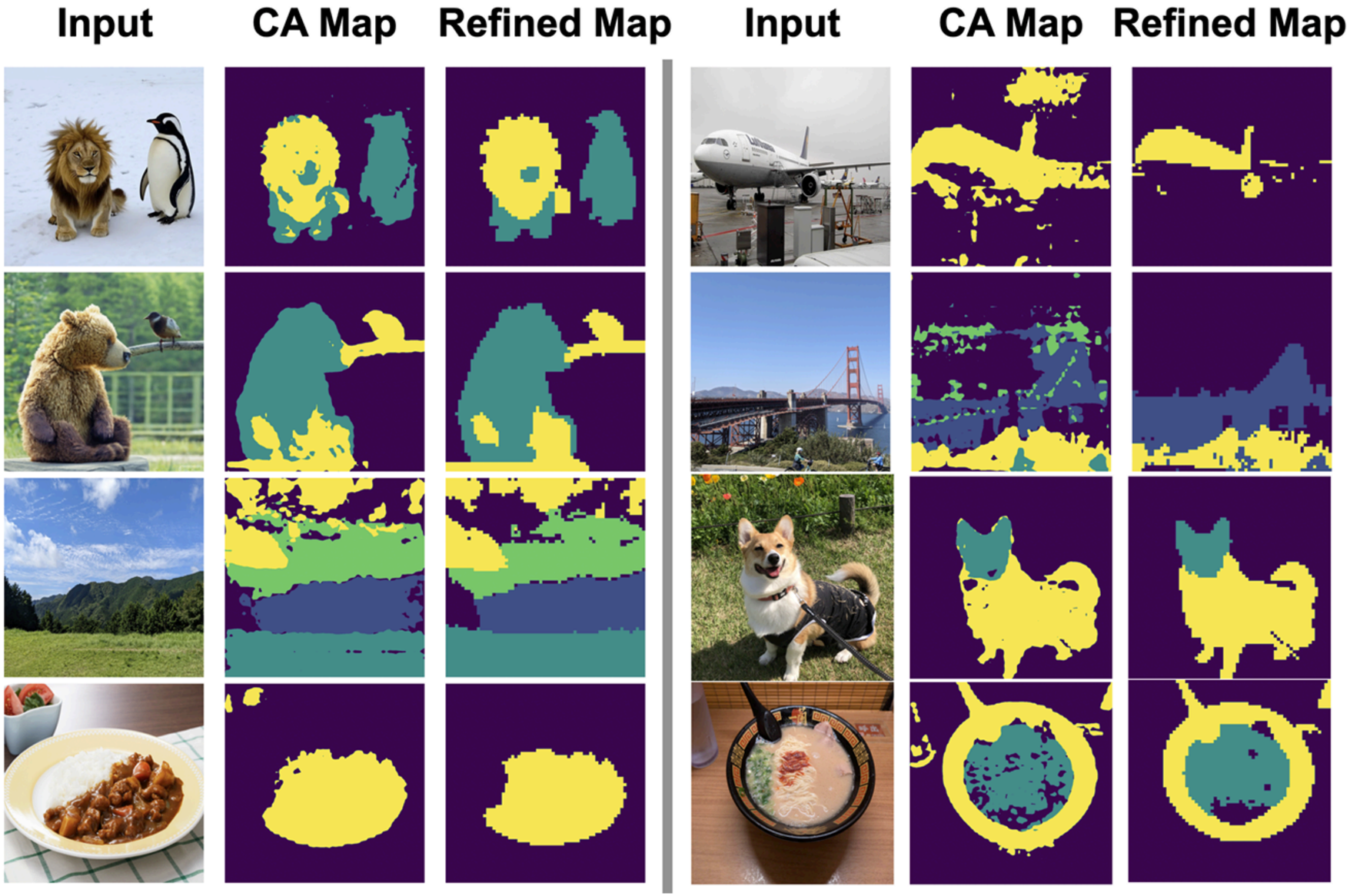
Appendix

How to build text embeddings




“a photo of ”
“aeroplane”
Tokenizer
(7,)
Input Ids
<bos>, a, photo, of, aero, plane, <eos>
Text Encoder
(7, 768)
Text Embedding
Sum
(1, 768)
“aeroplane” embedding

Additional Results

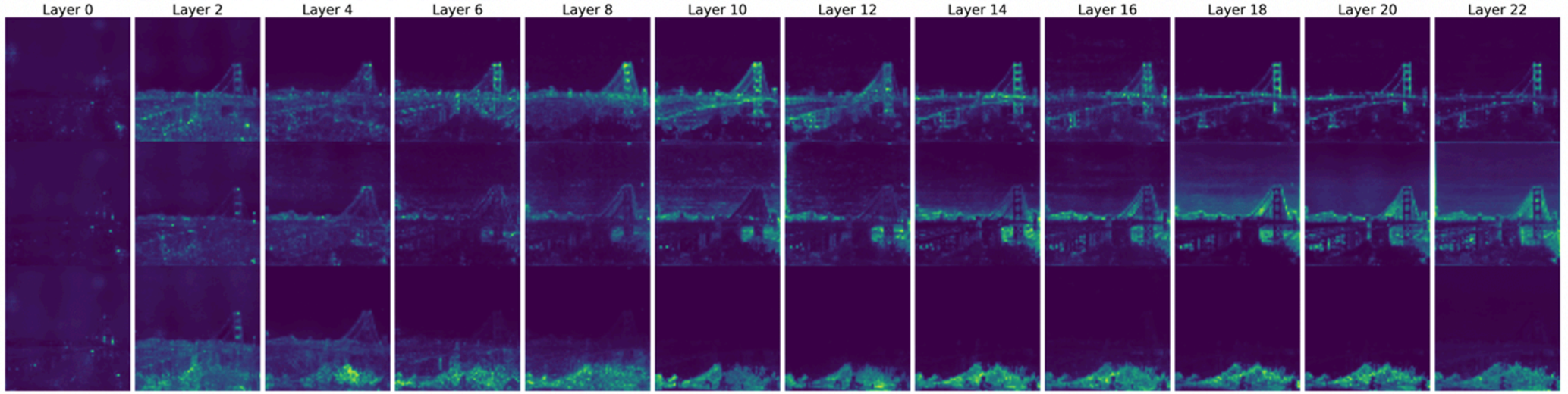


MM-DiT's Layer-wise CA Mask

Input Image



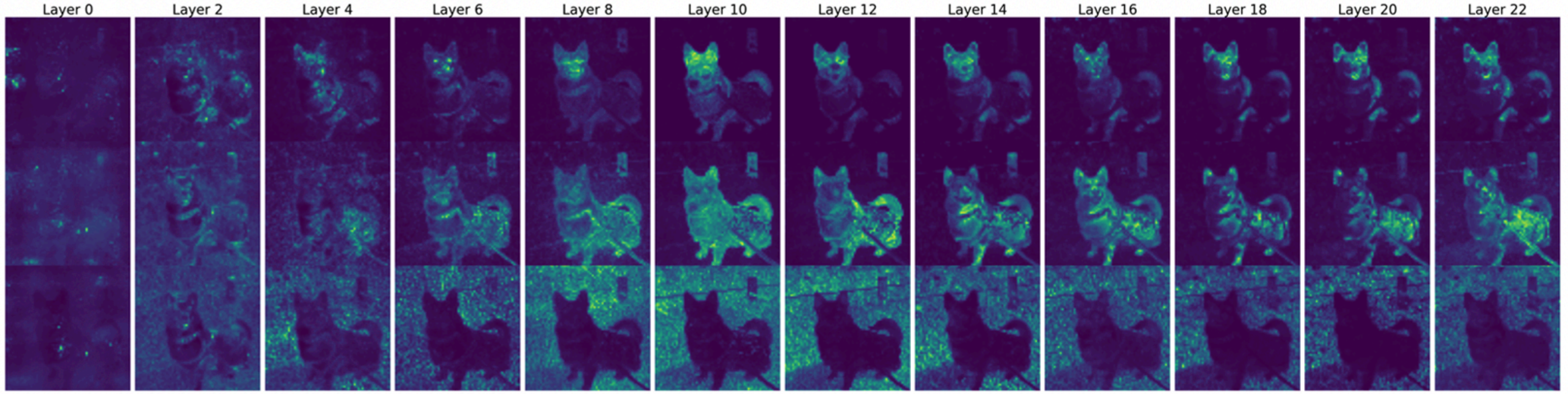
"Golden Gate"
"blue sky"
"trees"



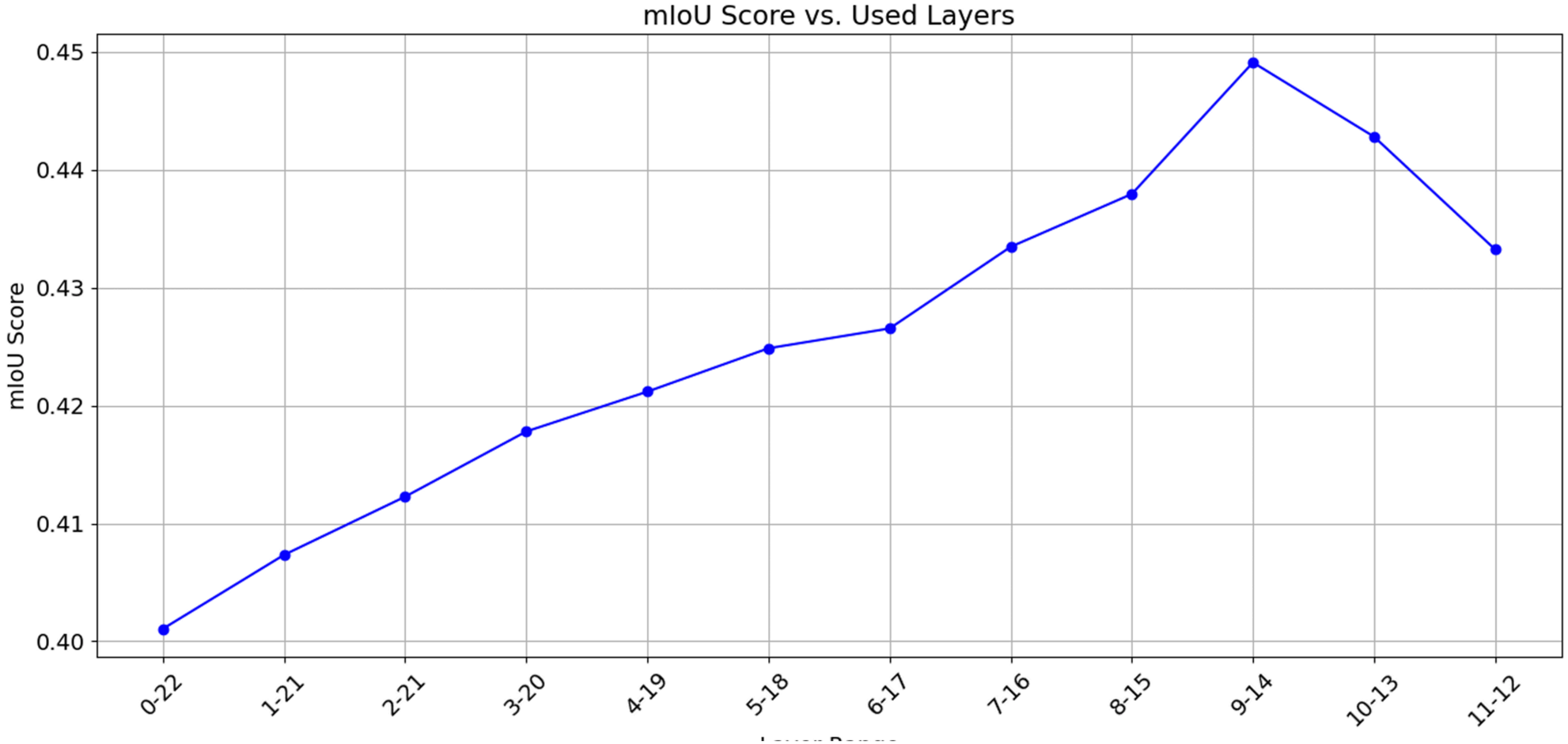
Input Image



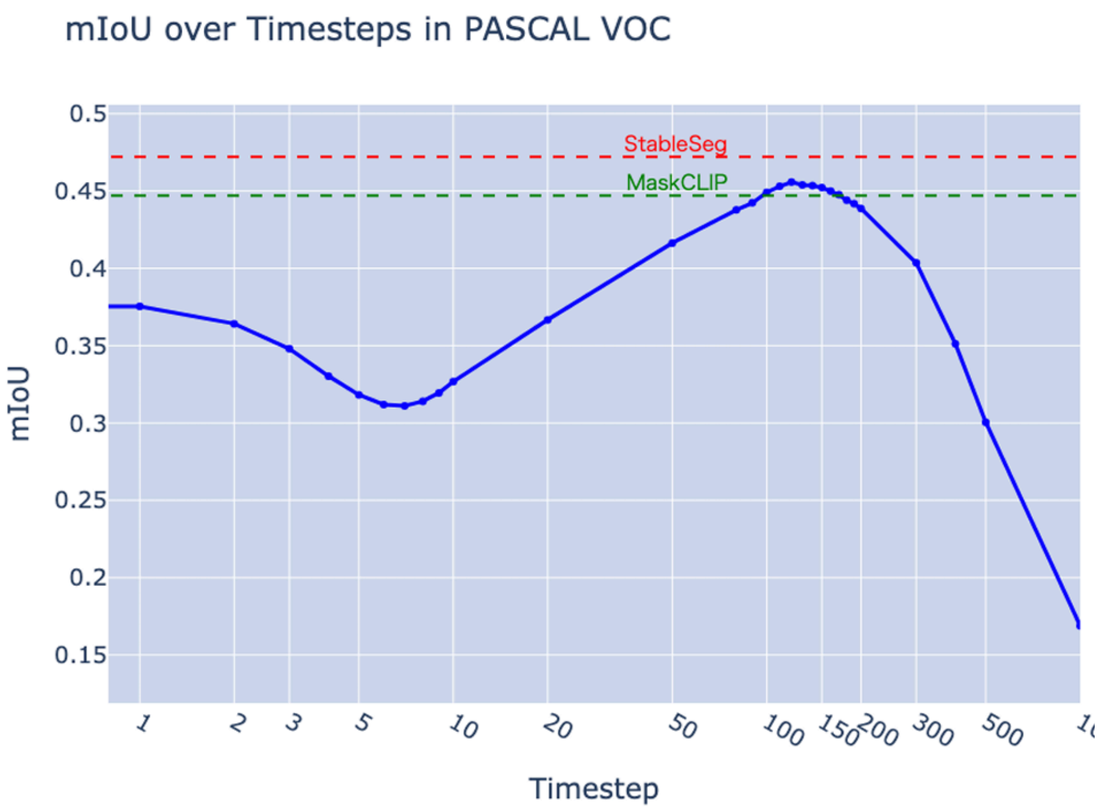
"corgi dog"
"clothes"
"grass"

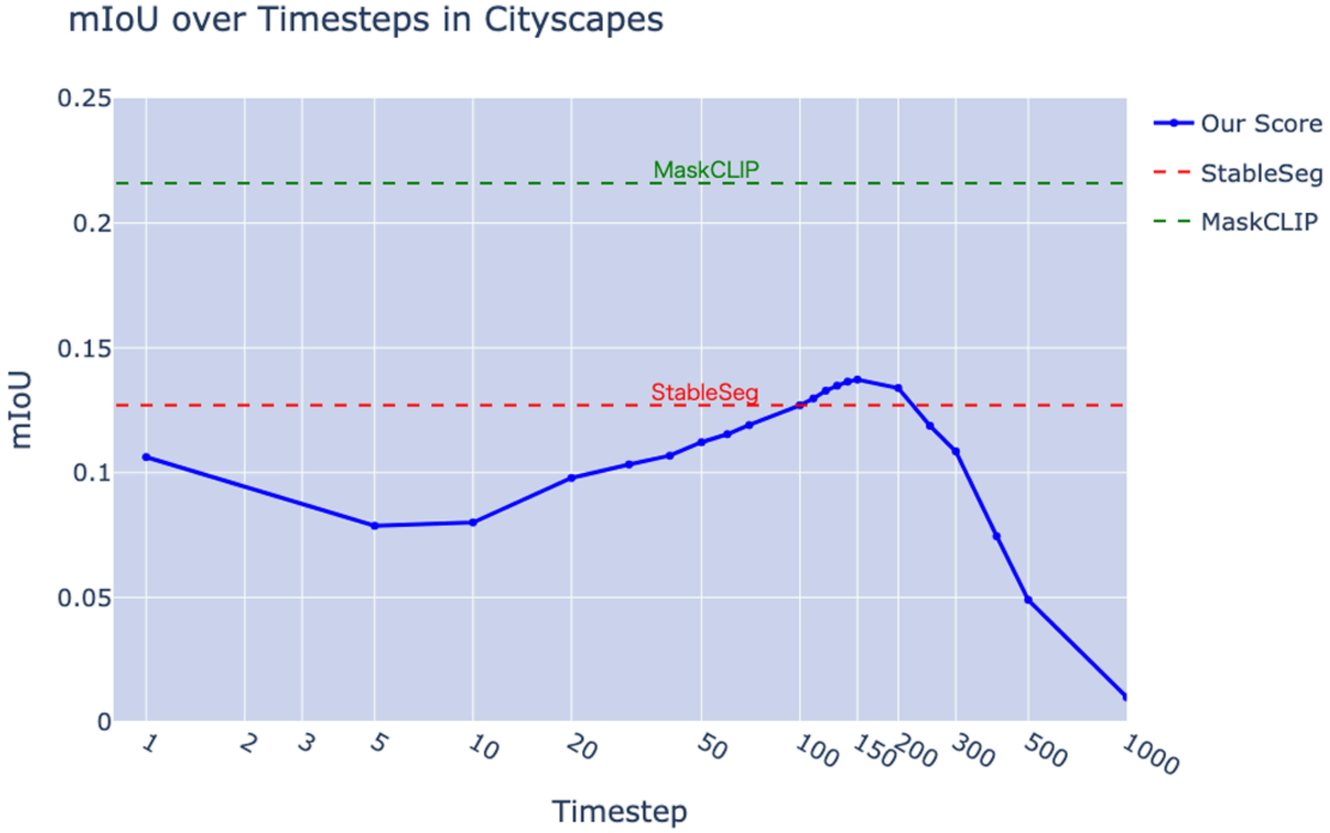


使用する MM-DIT の層とスコアの関係

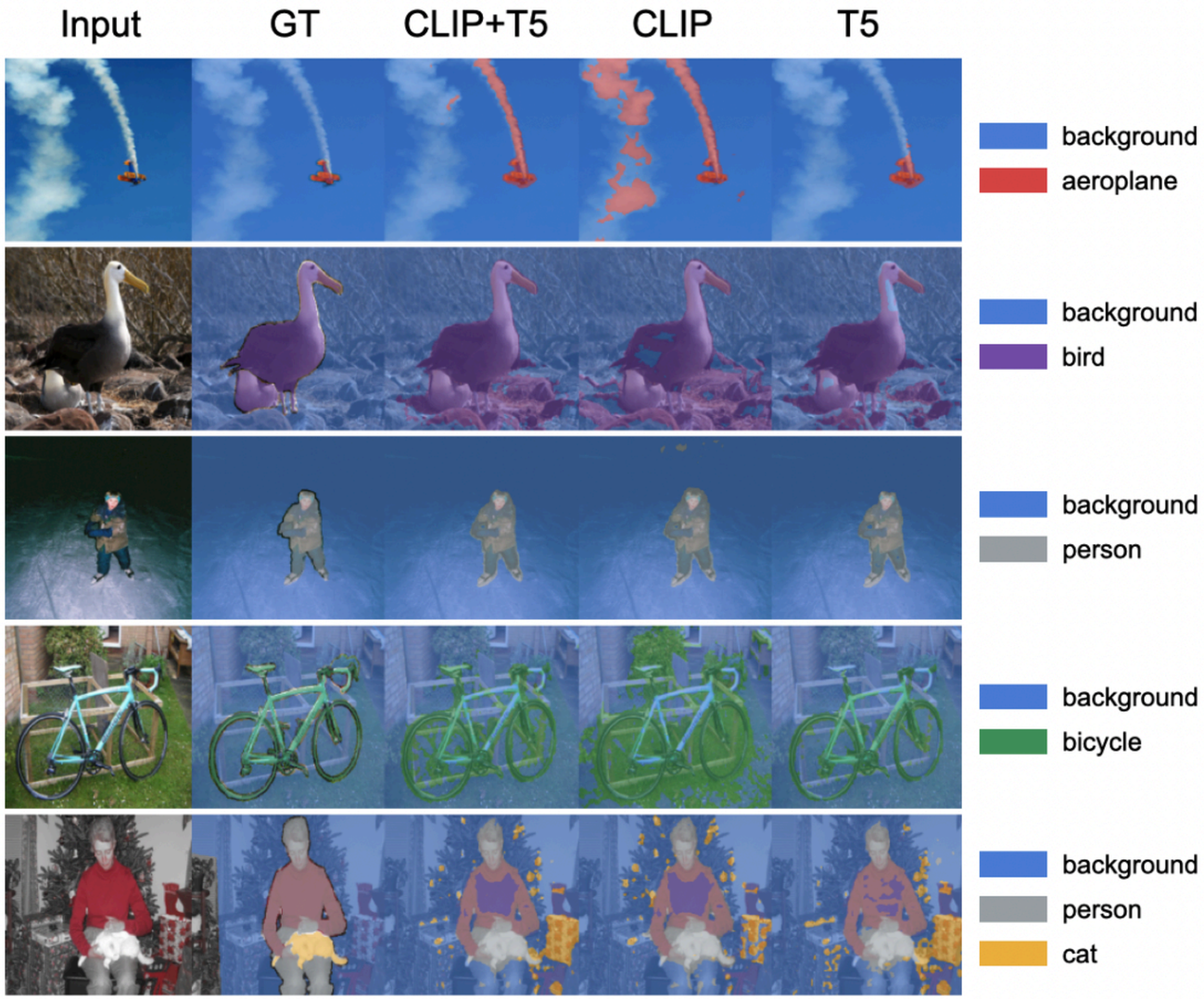


タイムステップとスコアの関係

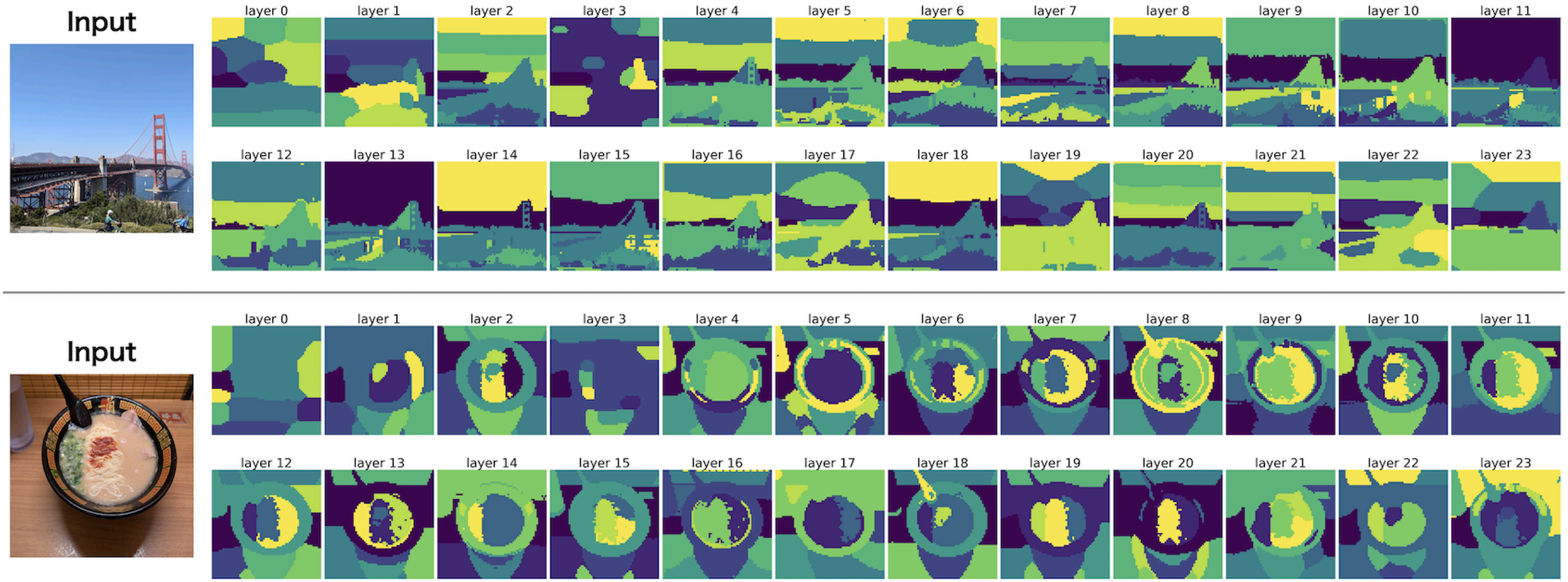




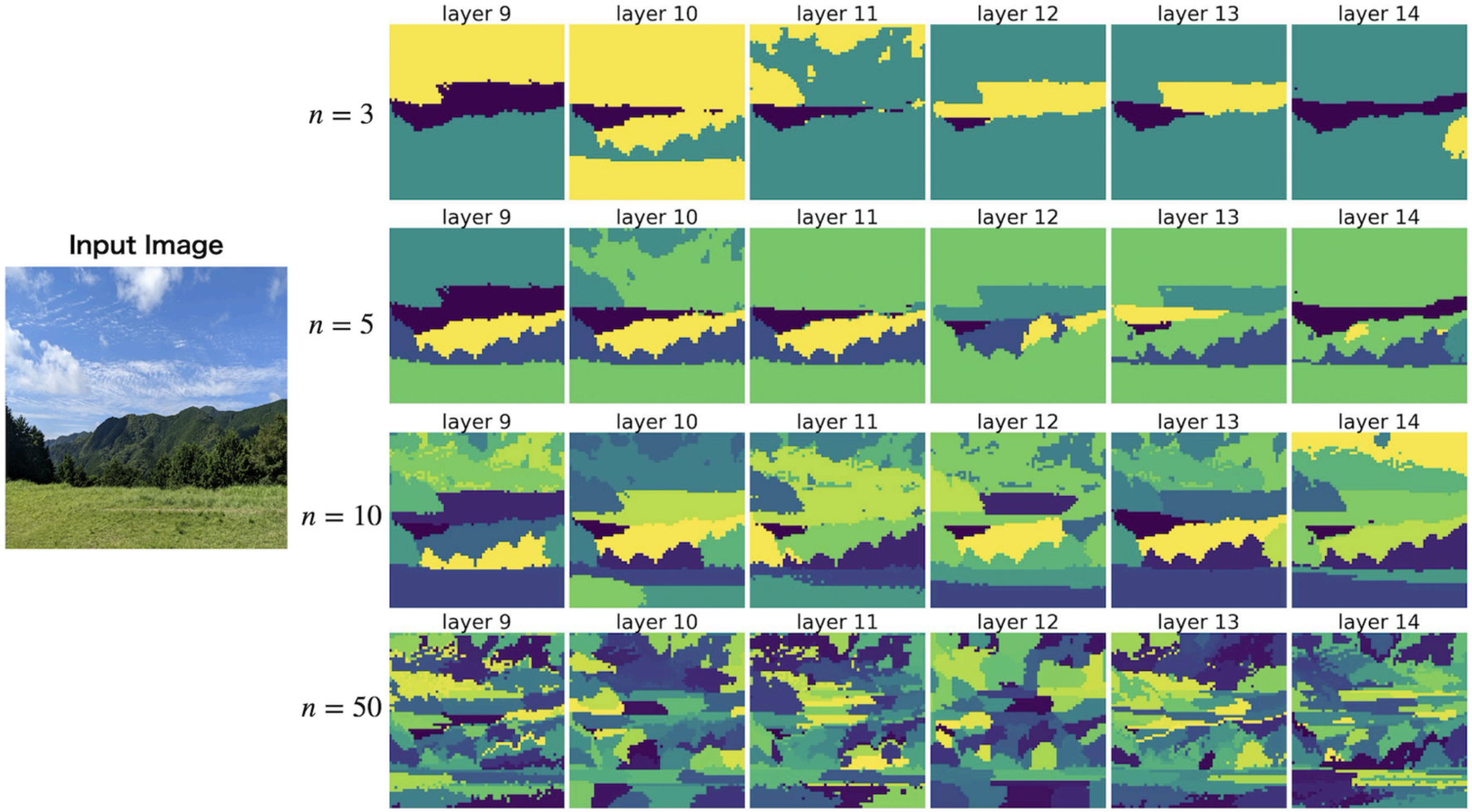
PACAL VOC におけるテキストエンコーダの違い



MM-DiT の層ごとのクラスタの変化



K-Means のクラスタ数によるSA Mask の変化



クラスタ数とスコアの関係

表 4.1: クラスタ数と mIoU の関係.

	10	**20**	**30**	**50**	**100**	**200**	**w/o SA**
mIoU	0.448	0.461	0.462	0.462	0.461	0.459	0.452