EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Textual Query-Driven Mask Transformer for Domain Generalized Segmentation

Byeonghyun Pak*   Byeongju Woo*   Sunghwan Kim*

Dae-hwan Kim   Hoseong Kim

https://byeonghyunpak.github.io/tqdm/

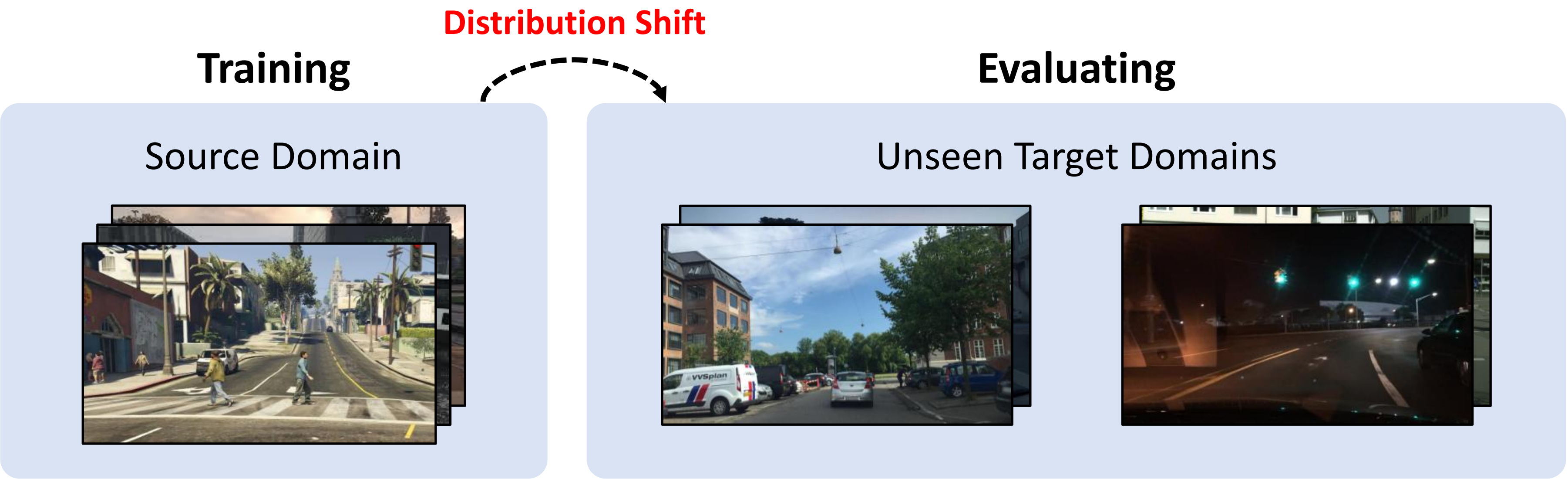
Agency for Defense Development

Problem Definition

Background

Domain Generalized Semantic Segmentation (DGSS)

- DGSS aims to build segmentation models that can generalize across unseen target domains, trained only on a single source domain.
    - *e.g.*, training on synthetic images (GTA5), testing on unseen real-world images (Cityscapes)

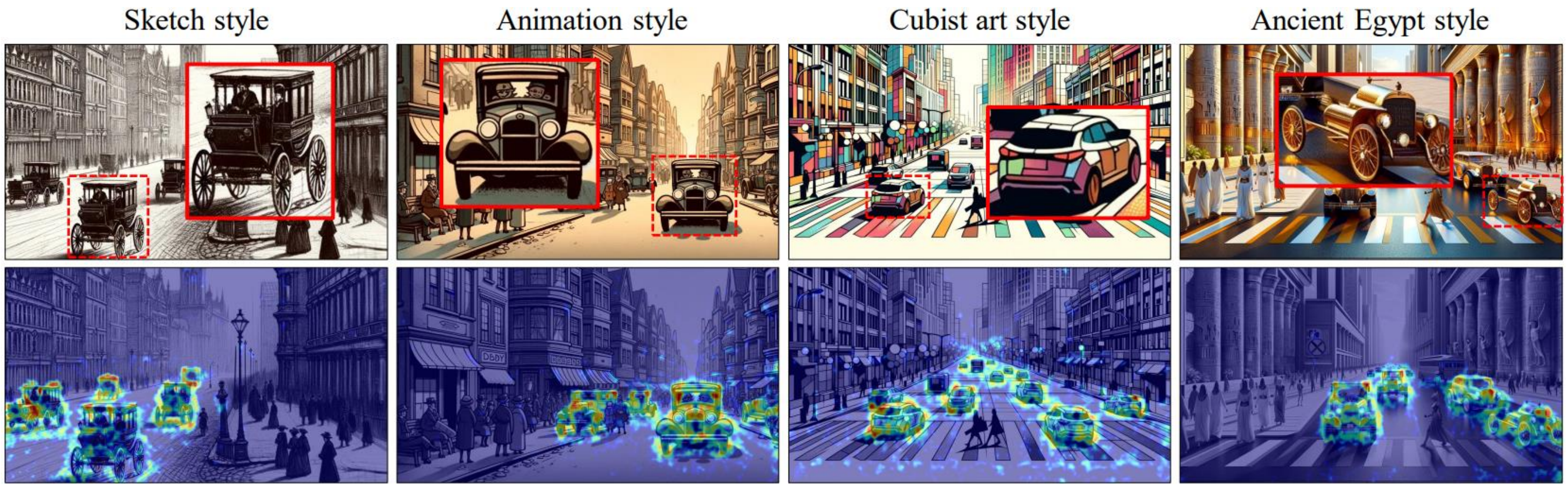


02 Motivation

Domain-Invariant Semantic Knowledge in VLMs

Our Key Observation:

- VLM's text embeddings encode domain-invariant semantic knowledge even at the pixel-level.
  - The semantic knowledge of text embeddings stems from web-scale contrastive learning objective of VLM.
  - The image-text similarity maps show strong activation in the corresponding regions across various domains.
  - ∴ One can leverage the textual information from VLMs for domain-generalized dense predictions.

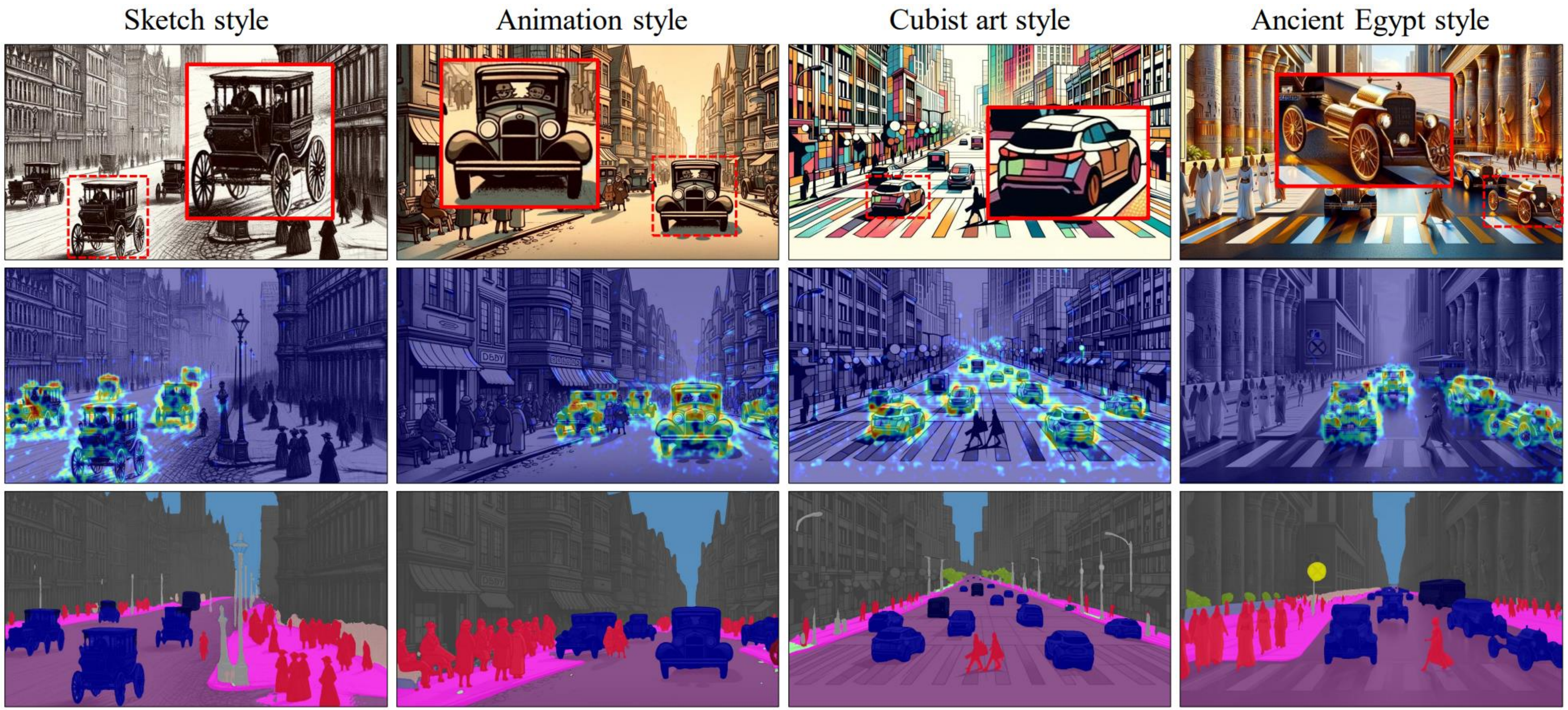


▲ Image-text similarity map of a pre-trained VLM. The text embedding of 'car' is consistently well-aligned with the corresponding class regions of images across various domains.

02

Motivation

Domain-Invariant Semantic Knowledge in VLMs



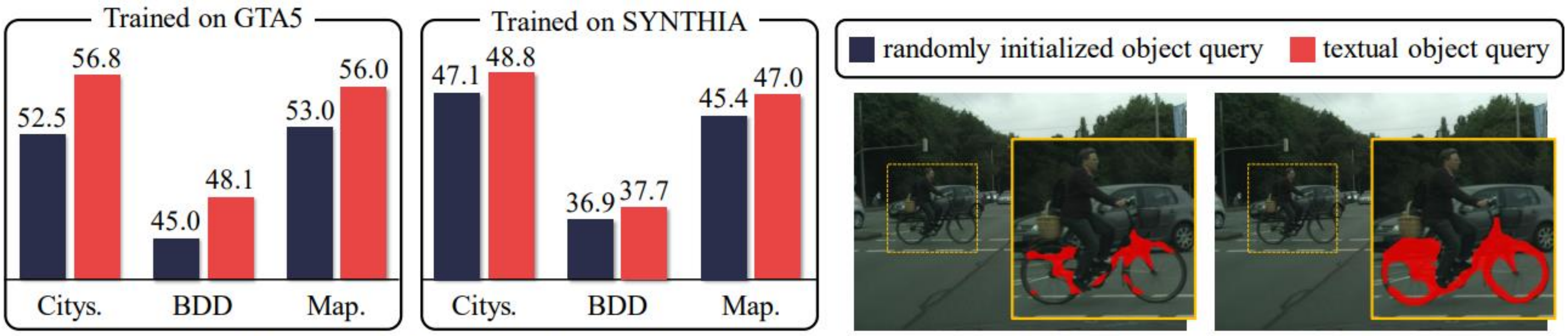
▲ Our proposed method can generalize to extreme domain shift, and effectively recognize the cars in various forms that are not present in the source domain.

Motivation

Domain-Invariant Semantic Knowledge in VLMs

Our Key Idea: *Textual Object Query*

- We propose a **textual object query-based segmentation framework**.
  - Utilizes the VLM's text embeddings of target classes as object queries, referred to as *textual object queries*.
  - Implementing object queries with the text embeddings results in domain-generalized mask predictions.



(a) Synthetic-to-Real DGSS Result, mIoU(%)

(b) Mask Prediction Results

02

Motivation

Domain-Invariant Semantic Knowledge in VLMs

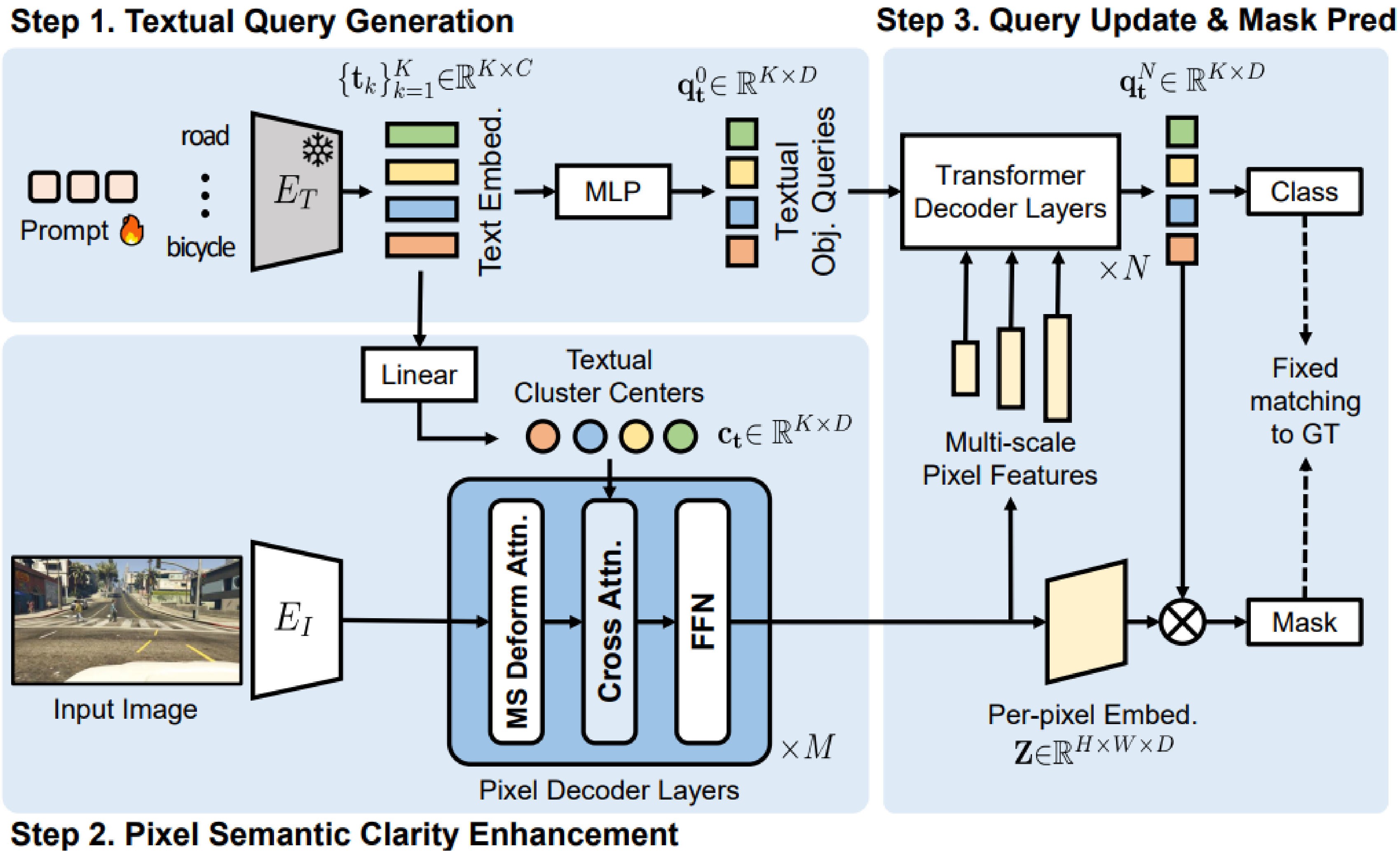
Our Key Idea: *Textual Object Query*

- We propose a **textual object query-based segmentation framework**.
  - Utilizes the VLM's text embeddings of target classes as object queries, referred to as *textual object queries*.
  - Implementing object queries with the text embeddings results in domain-generalized mask predictions.
- We design a **textual query-driven mask transformer (tqdm)** based on the following principles:
  1. Generate the object queries that maximally encode domain-invariant semantic knowledge.
  2. Improve the adaptability of queries for dense predictions by enhancing semantic clarity of pixel features.

03

Textual Query-Driven Mask Transformer

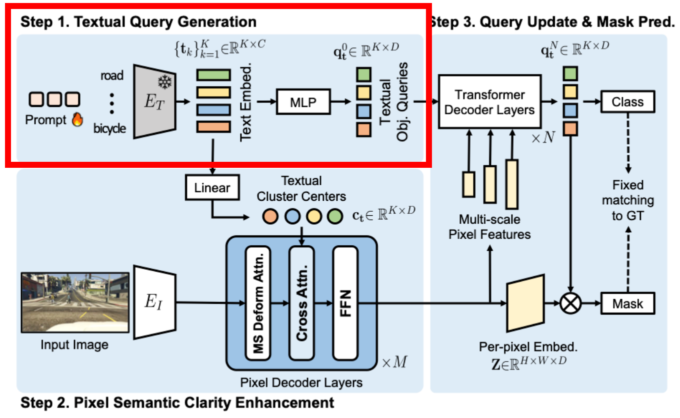
Overall Architecture



03

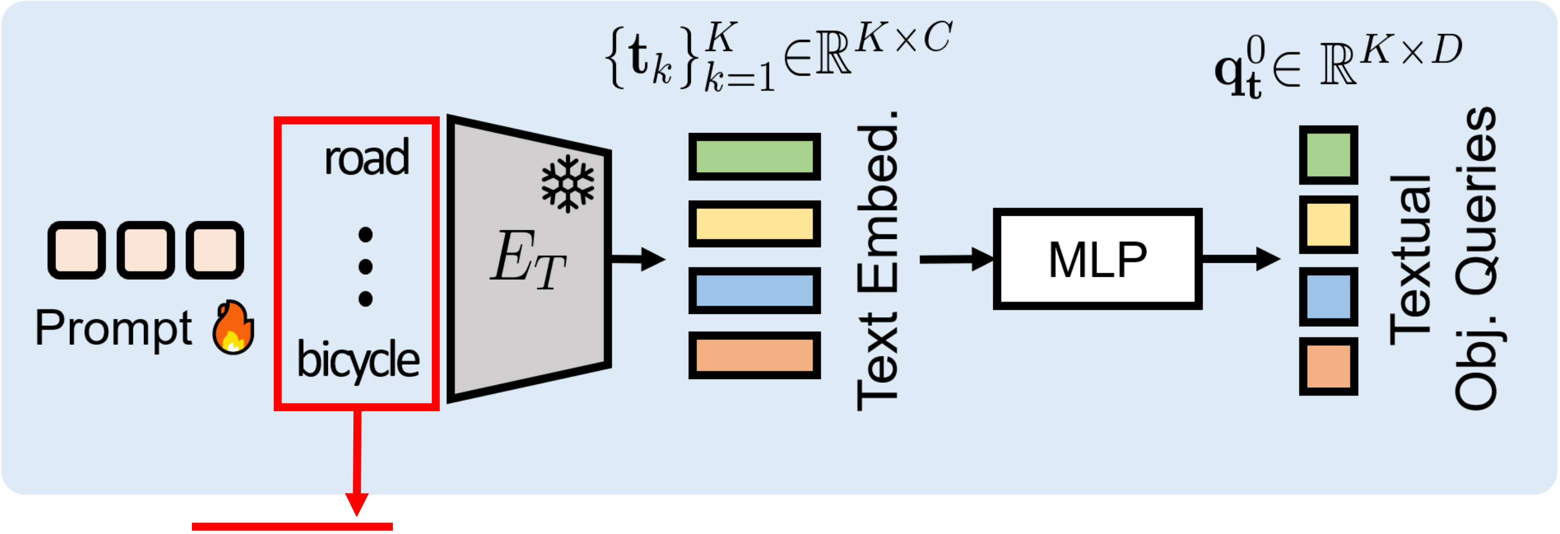
Textual Query-Driven Mask Transformer

Step 1. Textual Query Generation



- Generate initial textual object queries $\mathbf{q}_t^0$ from K text embeddings $\{\mathbf{t}_k\}_{k=1}^{K}$.
  - Each textual query encode the semantic of the corresponding class.
- The object queries should …
  - (1) preserve domain-invariant semantics ➡ freeze the pre-trained text encoder of VLM.
  - (2) adapt to the segmentation task ➡ employ the learnable prompt.

Step 1. Textual Query Generation

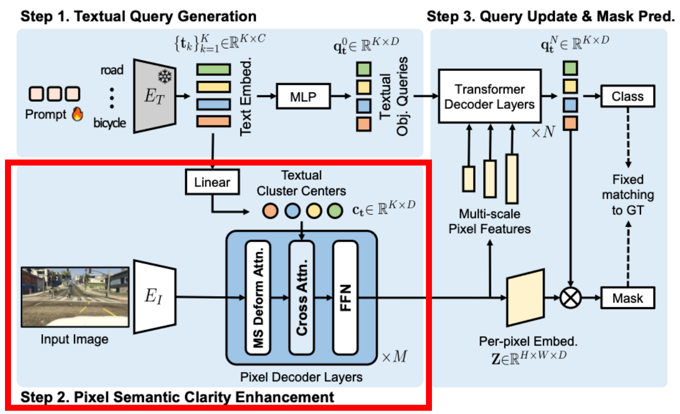


Each class name among the K = 19 classes ➡ 19 textual object queries

03

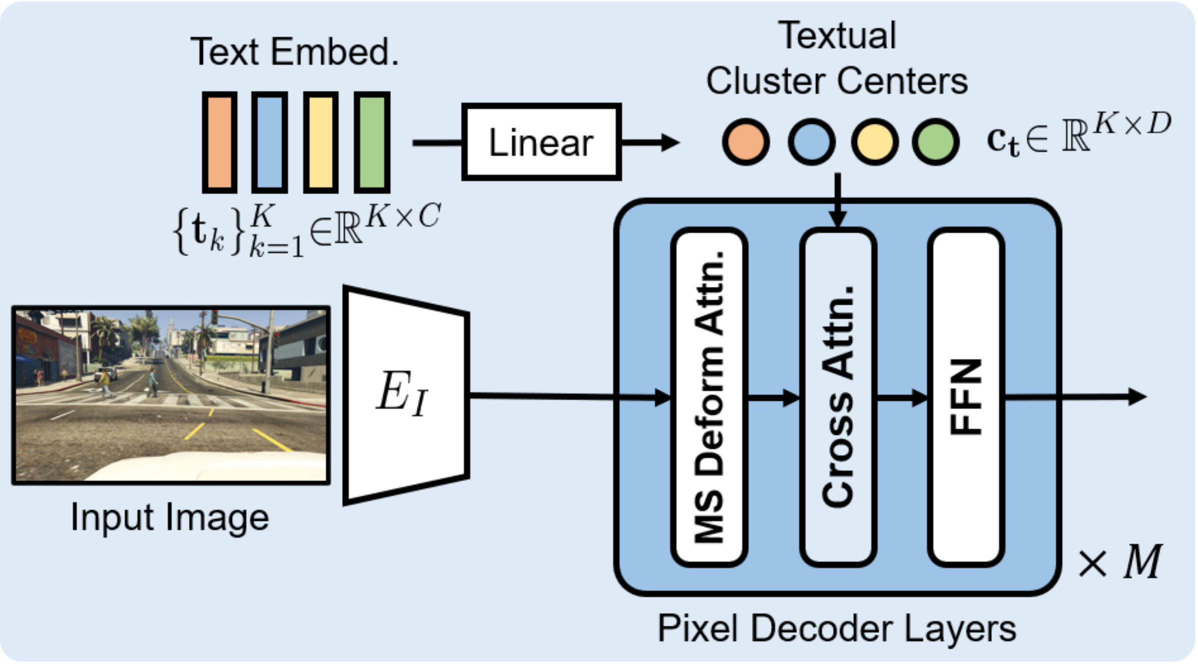
Textual Query-Driven Mask Transformer

Step 2. Pixel Semantic Clarity Enhancement

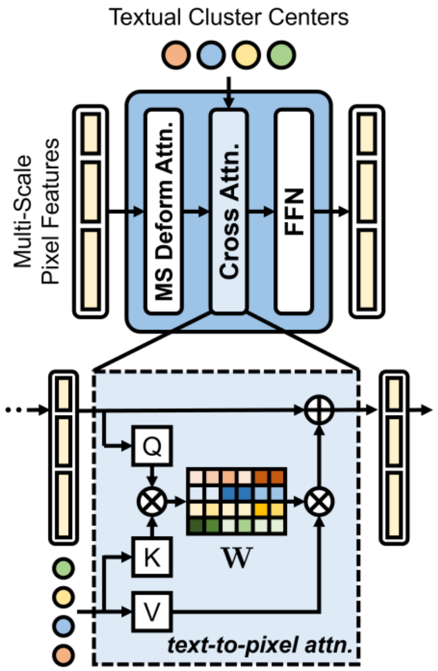


- We incorporate <u>*text-to-pixel attention*</u> within the pixel decoder.
  - Aligns the pixel features with the corresponding textual cluster centers.
  - Enhances the pixel semantic clarity.

  ➡ Ensures that pixel features are clearly represented in terms of domain-invariant semantics.
  - Allows the pixel features to be effectively grouped by textual object queries.



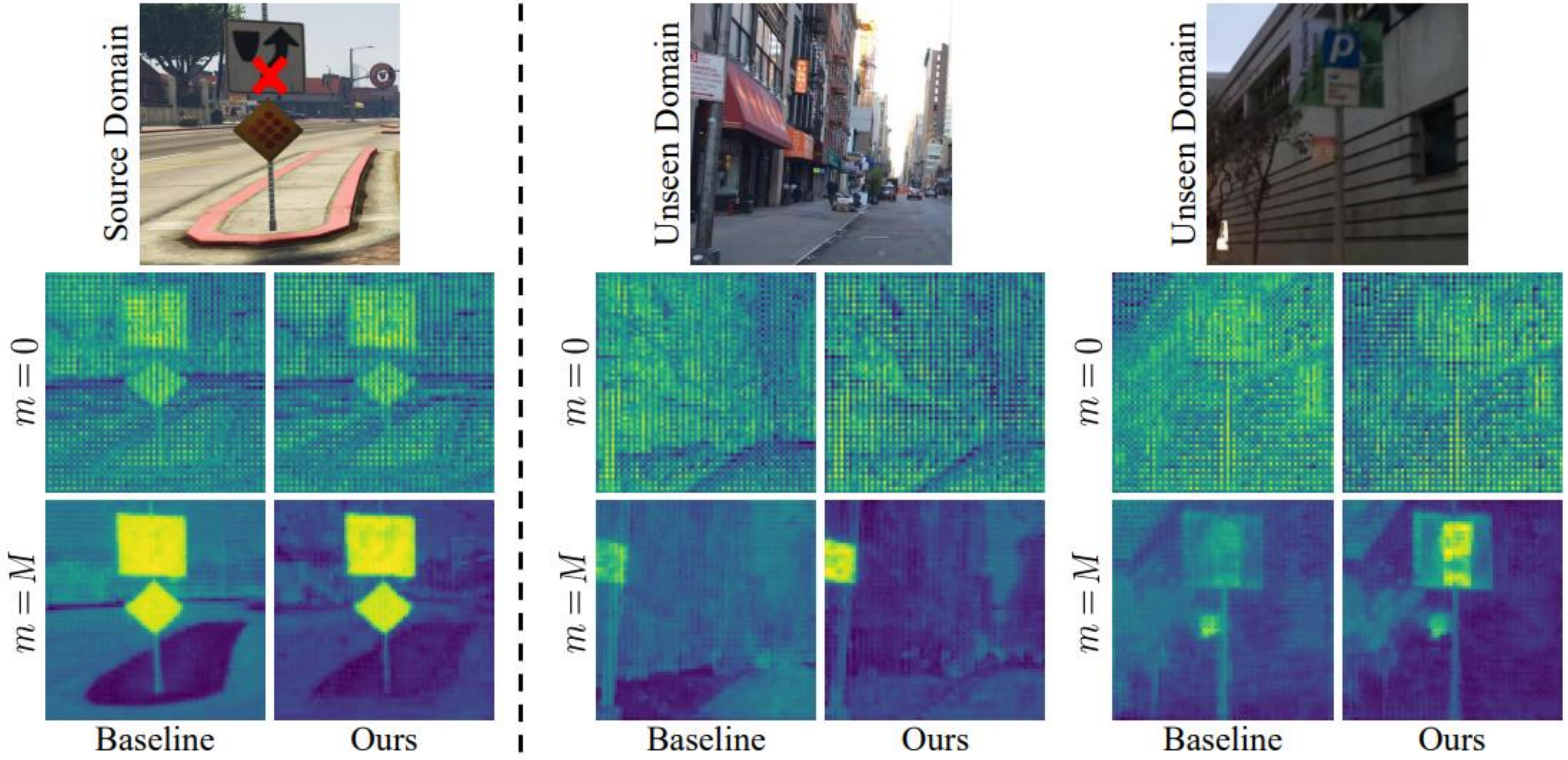
Step 2. Pixel Semantic Clarity Enhancement



03 Textual Query-Driven Mask Transformer

Step 2. Pixel Semantic Clarity Enhancement

Ablation Result on Pixel Semantic Clarity



$m = 0$: before the pixel decoder

$m = M$: after the pixel decoder

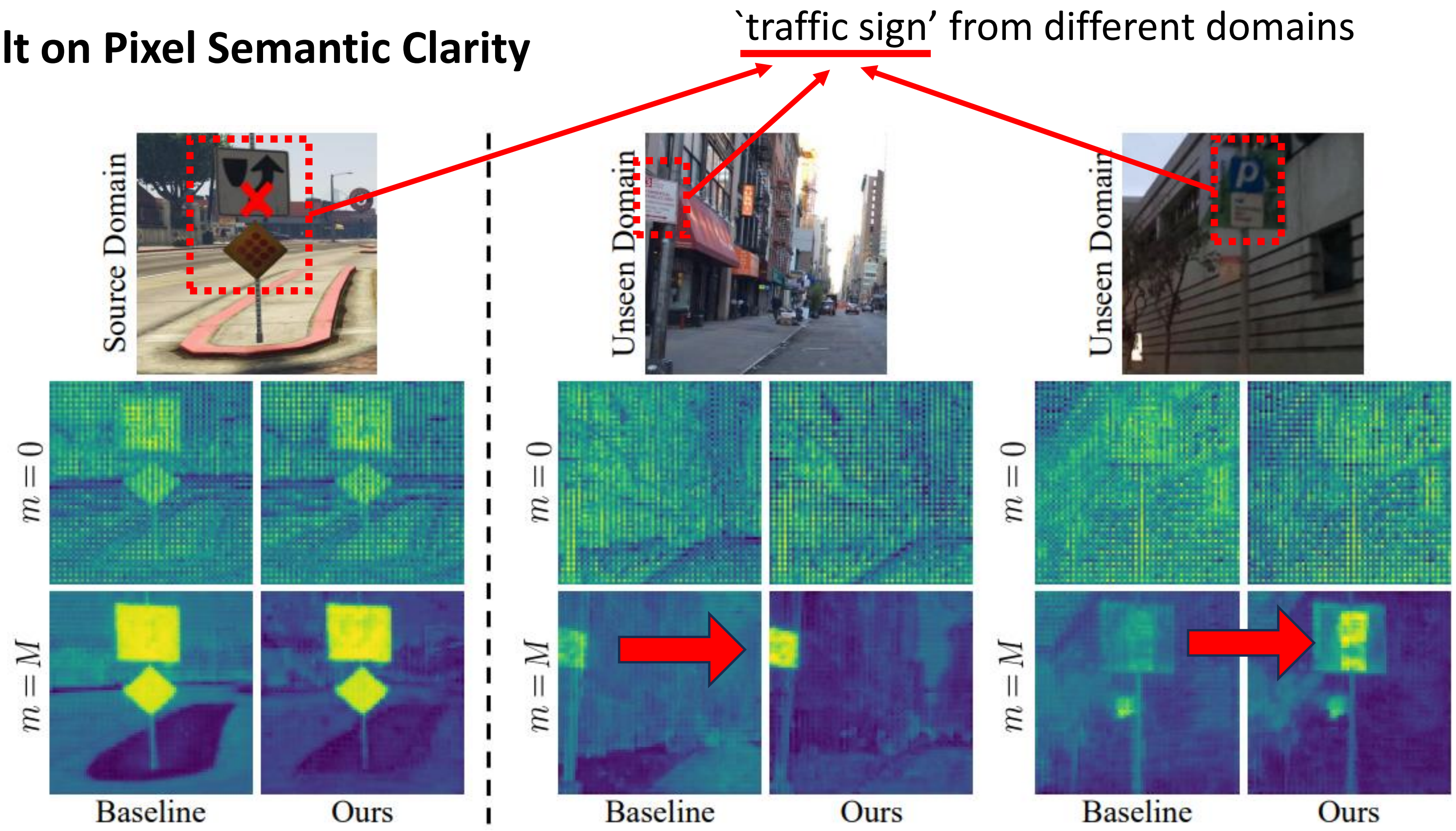
Baseline: w/o text-to-pixel attention

Ours: w/ text-to-pixel attention

03 Textual Query-Driven Mask Transformer

Step 2. Pixel Semantic Clarity Enhancement

Ablation Result on Pixel Semantic Clarity



$m = 0$: before the pixel decoder

$m = M$: after the pixel decoder

Baseline: w/o text-to-pixel attention

Ours: w/ text-to-pixel attention

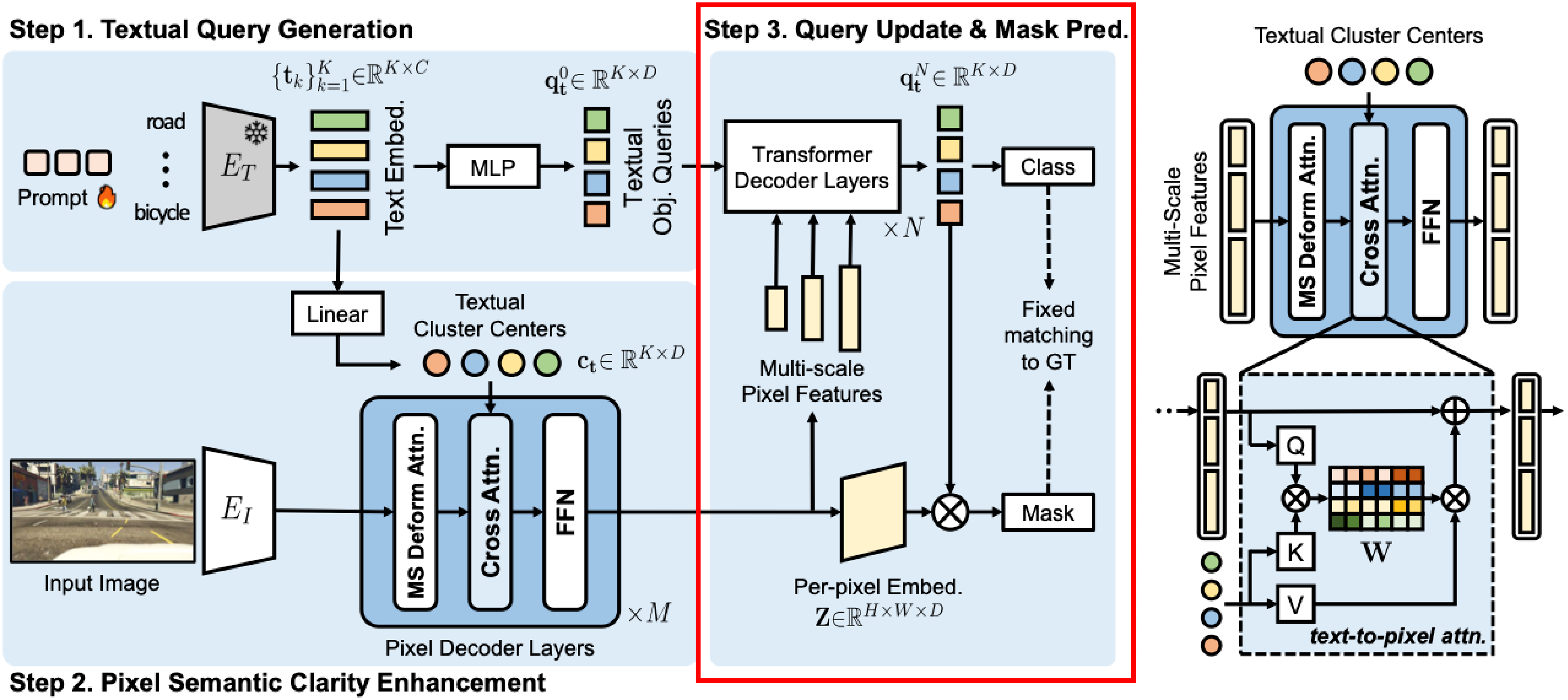
03

Textual Query-Driven Mask Transformer

Step 3. Query Update & Mask Prediction

Objective Loss: $\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{reg}$

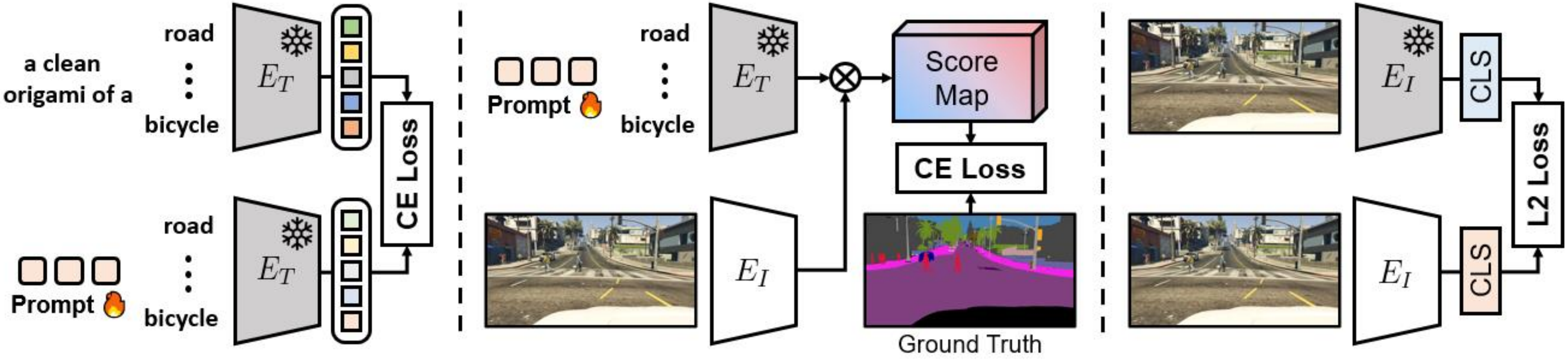
- $\mathcal{L}_{seg} = \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice} + \lambda_{cls}\mathcal{L}_{cls}$
- $\mathcal{L}_{reg} = \mathcal{L}_{reg}^{L} + \mathcal{L}_{reg}^{VL} + \mathcal{L}_{reg}^{V}$



03

Textual Query-Driven Mask Transformer

Regularization Loss



(a) Language Regularization $\mathcal{L}_{reg}^{L}$ (b) Vision-Language Regularization $\mathcal{L}_{reg}^{VL}$ (c) Vision Regularization $\mathcal{L}_{reg}^{V}$

(a) $\mathcal{L}_{reg}^{L}$ prevents the learnable prompt from distorting the semantic meaning of the text embeddings.

(b) $\mathcal{L}_{reg}^{VL}$ preserves joint vision-language alignment at the pixel-level.

(c) $\mathcal{L}_{reg}^{V}$ maintains the visual backbone’s alignment with the text embeddings at the image-level.

04 Experimental Results

Quantitative Results

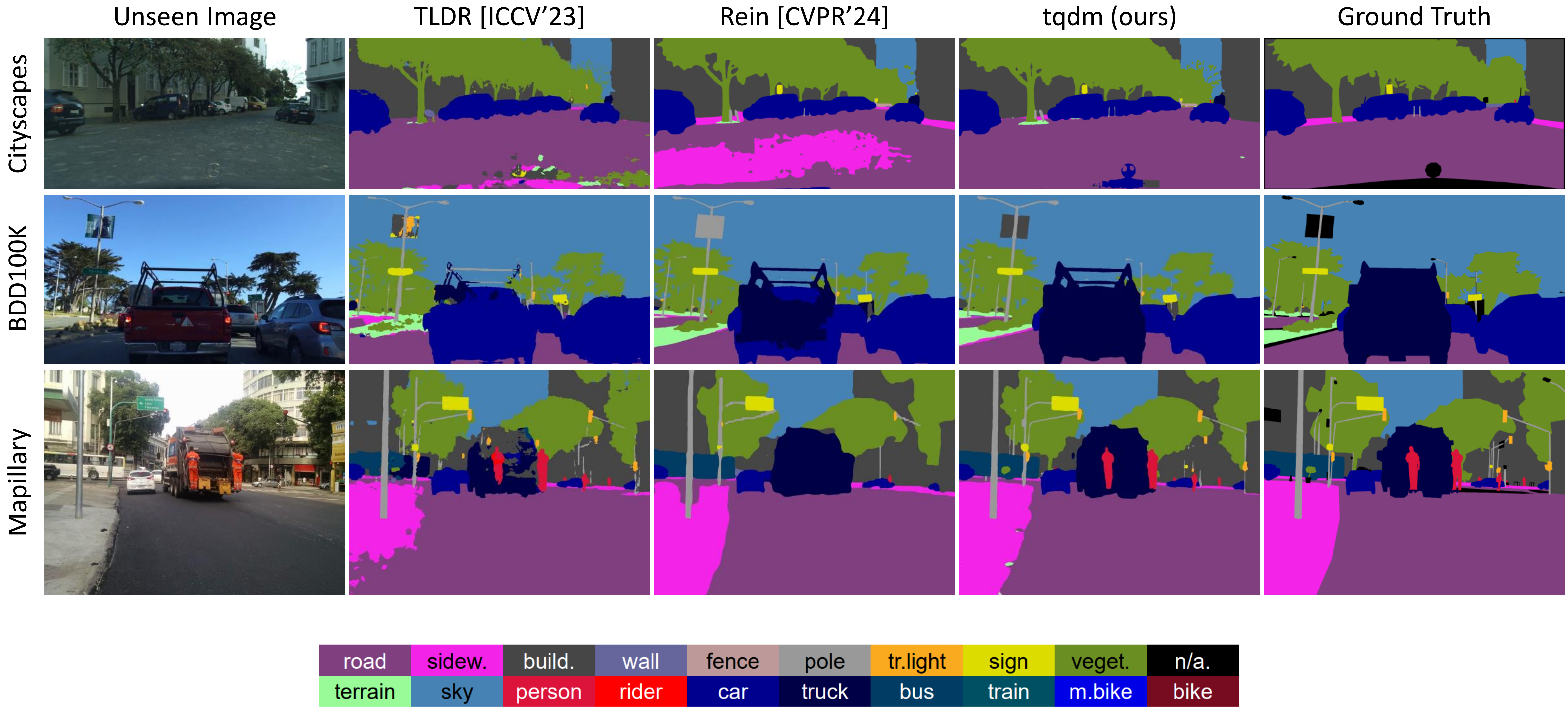
Method	Backbone	*synthetic-to-real* G→C	G→B	G→M	**Avg.**	*real-to-real* C→B	C→M	**Avg.**
SHADE [64]	MiT-B5	53.27	48.19	54.99	52.15	-	-	-
IBAFormer [50]	MiT-B5	56.34	**49.76**	58.26	54.79	-	-	-
VLTSeg [20] 📎	ViT-B	47.50	45.70	54.30	49.17	-	-	-
`tqdm` (ours) 📎	ViT-B	**57.50**	47.66	**59.76**	**54.97**	**50.54**	**65.74**	**58.14**
HGFormer [10]	Swin-L	-	-	-	-	61.50	72.10	66.80
VLTSeg [20] 🖇️	EVA02-L	65.60	58.40	66.50	63.50	64.40†	**76.40**†	70.40†
Rein [54] 🖇️	EVA02-L	65.30	**60.50**	64.90	63.60	64.10	69.50	66.80
Rein [54] 🦖	ViT-L	66.40	60.40	66.10	64.30	**65.00**	72.30	68.65
`tqdm` (ours) 🖇️	EVA02-L	**68.88**	59.18	**70.10**	**66.05**	64.72	76.15	**70.44**

📎: CLIP 🖇️: EVA-CLIP 🦖: DINOv2

The best and second-best results are **highlighted** and underlined, respectively

04 Experimental Results

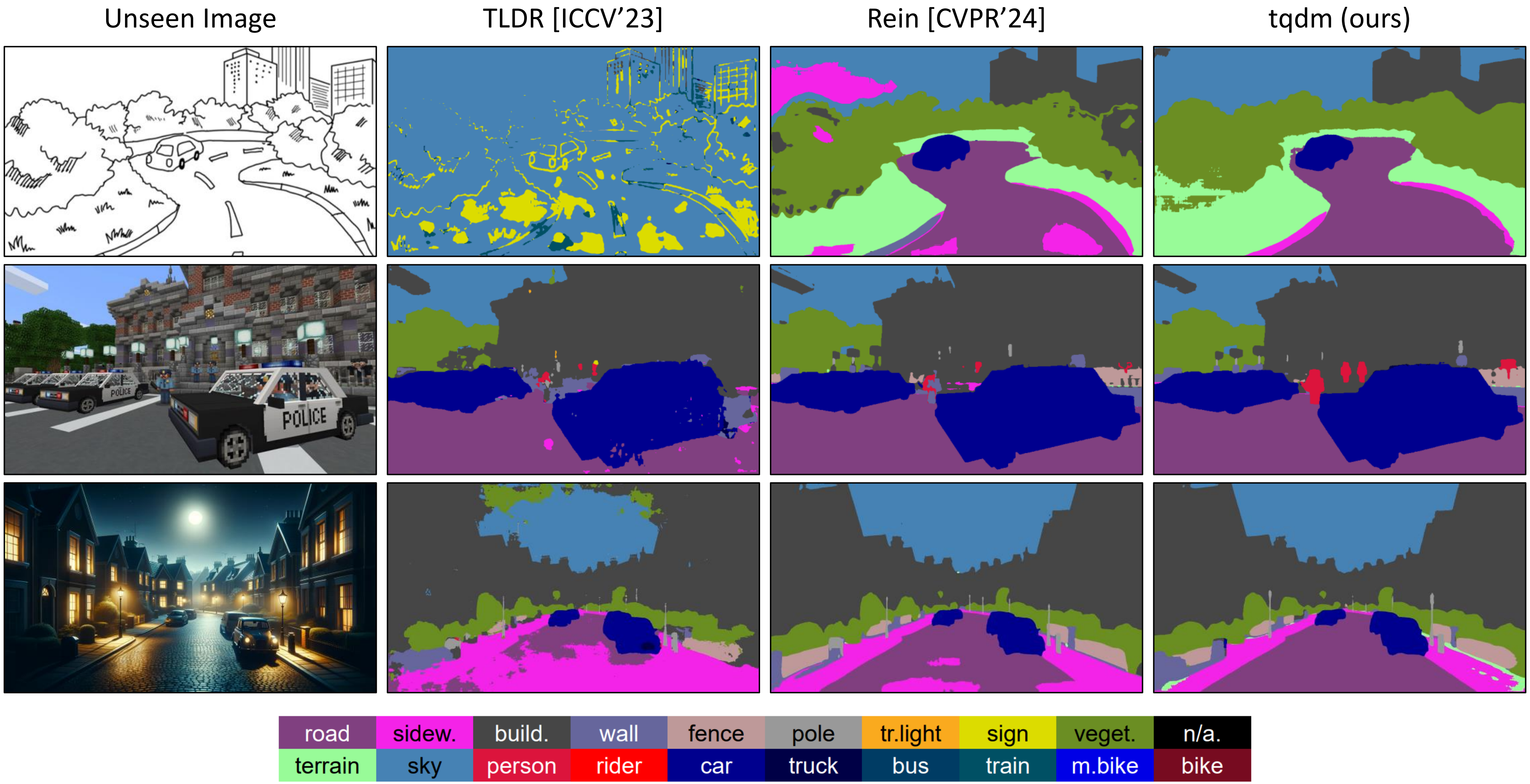
Qualitative Results on Benchmarks



04

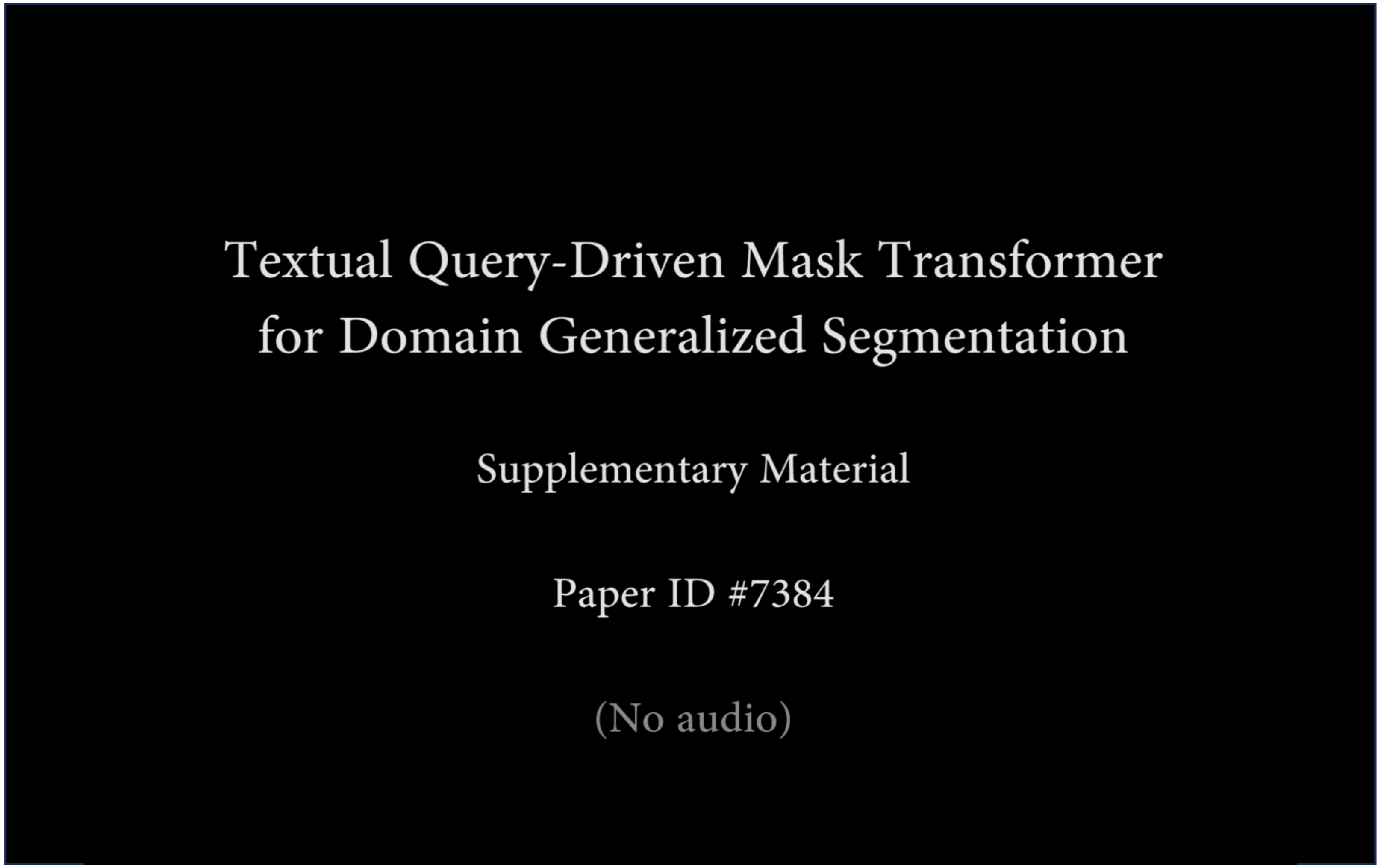
Experimental Results

Qualitative Results under Extreme Domain Shifts



Experimental Results

Qualitative Results under Extreme Domain Shifts



road	sidew.	build.	wall	fence	pole	tr.light	sign	veget.	n/a.
terrain	sky	person	rider	car	truck	bus	train	m.bike	bike

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Thank you for your attention

https://byeonghyunpak.github.io/tqdm/

Code and paper are available here.