Hierarchical Prompt Injector for Domain Generalization Segmentation

Xin Kun Lin*, Ruoyu Guo*, Jiaqi Guo, Maurice Pagnucco, and Yang Song

School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
{xin_kun.lin,morri,yang.song1}@unsw.edu.au
{ruoyu,jiaqi}@student.unsw.edu.au

Abstract. Domain Generalized Semantic Segmentation (DGSS) is a challenging task, as vision models often rely on low-level appearance cues that change across domains. In contrast, structural attributes exhibit cross-domain stability, motivating the use of structural priors for DGSS. Existing methods use prompt learning to transfer such priors into DGSS models, but typically encode each class as a single holistic prompt. Moreover, these methods apply prompts uniformly to all pixels, offering no mechanism to adapt when only a subset of object regions is visible due to viewpoint changes, occlusion, and environmental variation. We address this with **Spatial Hierarchical Prompts (SHP)** that enrich each class with region-level geometric anchors capturing structural appearance from distinct viewing angles, ensuring complementary coverage under arbitrary viewpoints. Additionally, we propose the **Hierarchical Prompt Injector (HPI)**, which enables spatially adaptive prompt injection in foundation models. HPI spatially grounds prompts by modeling their semantic relevance and spatial influence with visual features. Considering the difficulty of learning spatially and semantically aware prompt injection, we further introduce auxiliary supervision to align hierarchical prompts with their corresponding object regions. We achieve 70.62% and 72.74% mIoU on synthetic-to-real and real-to-real benchmarks, respectively. Code and checkpoints are released at `https://github.com/MosukFate/HPI`.

Keywords: Domain Generalization · Semantic Segmentation · Vision-Language Models · Prompt Learning

1 Introduction

Domain generalized semantic segmentation (DGSS) aims to learn segmentation models from labeled source domains that can generalize to unseen target domains, where no target data are available during training [39]. Classical DGSS

* These authors contributed equally.

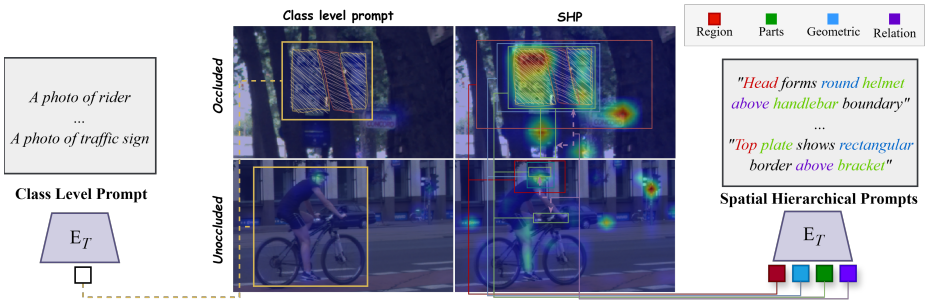

Fig. 1: Class-level prompts (left) versus our Spatial Hierarchical Prompts (right). Class-level prompts encode each category as a single holistic sentence, producing diffuse attention that spreads across background regions. SHP decomposes each class into region-level prompts built from four complementary components: Region Cue, Part, Geometry, and Boundary Relation.

strategies, such as style augmentation [61, 64] and frequency-domain perturbation [6, 31, 47], seek to reduce the model's reliance on appearance cues, but operate solely on low-level statistics without exploiting the underlying geometric structure of objects. Yet recent studies [12, 20, 37] have shown that structural attributes exhibit strong cross-domain stability, offering more reliable anchors for pixel-wise recognition. Accordingly, prompt learning in Vision-Language Models (VLMs) such as CLIP [38] has become a mainstream strategy for learning such structural attributes [24, 26, 65, 66].

Recent prompt-learning methods construct textual prompts, such as class names [35, 40], short domain descriptors [19, 23], or learnable context tokens [65, 66], to guide VLM features toward domain-robust representations. However, these methods encode each category as a single holistic sentence, compressing the geometric diversity of a class into one embedding and discarding region-specific structural cues that remain stable across domains. As illustrated in Fig. 1, such class-level prompts produce diffuse attention that fails to distinguish object parts from the background, and the problem is amplified under occlusion where a single description cannot match the limited visible content. Moreover, existing methods either feed prompts only into the decoder without modifying backbone features [19,35], or inject them into the encoder uniformly across all positions [4,40], with no mechanism to control *where* prompt information influences dense visual features. Most also lack explicit supervision over prompt-to-region correspondences [5, 12, 48], leaving the spatial grounding of prompts uncontrolled. This spatial imprecision is further amplified by VLMs pretrained with image-level contrastive objectives [38], which inherently lack the ability to ground language onto individual patches [5, 23, 35]. These limitations raise a core question: *how can we encode stable geometric cues at the region level and inject them adaptively when only parts of an object are visible?*

We propose a VLM-based framework for DGSS that integrates region-level geometric prompts into visual representations in a spatially adaptive and content-

aware manner. We realize this framework with **Spatial Hierarchical Prompts (SHP)** and the **Hierarchical Prompt Injector (HPI)**. Specifically, SHP decomposes each class into a small set of region-level prompts from distinct viewing angles (*e.g.*, front, side, and rear of a vehicle), each encoding complementary geometric cues, including discriminative parts, shape attributes, and boundary relations. SHP provides stable geometric anchors that remain discriminative under distribution shift, because each regional description is tied to a semantically meaningful object part and emphasizes geometric structure and boundary relations, which are typically more stable across domains than low-level appearance. To make these textual anchors effective for dense prediction, HPI converts SHP embeddings into patch-wise feature residuals and injects them into visual features in a spatially adaptive manner. Since VLMs lack the spatial precision to ground prompts onto individual patches [5,35], we complement the VLM stream with a DINOv2 [34] spatial grounding branch, allowing HPI to inject semantic prompts with finer spatial support.

Within this two-branch injection framework, HPI controls not only *whether* a prompt is active at a given position, but also *how strongly* it modulates the features there. This is achieved through two complementary confidence maps. *Semantic Alignment Confidence* (SAC) measures the content-level relevance between each prompt and patch, determining *whether* a prompt should be active at a given position. *Spatial Coherence Confidence* (SCC) captures structural consistency across neighboring patches via a learned spatial confidence map, determining *how much* an active prompt modulates features while enforcing spatial coherence. We further introduce a *Spatial Localization Loss* to localize each region-level prompt to its corresponding object region, and a *Semantic Consistency Loss* to suppress responses from classes absent in the image. Our contributions are summarized as follows:

- We propose **Spatial Hierarchical Prompts** that decompose each class into region-level geometric anchors from distinct viewing angles. These prompts convert domain-invariant geometric structure into textual priors for pixel-level prediction under domain shifts.
- We develop the **Hierarchical Prompt Injector**, which adaptively controls *whether* and *how much* region-level prompts take effect via SAC and SCC. Our auxiliary losses, *Spatial Localization* and *Semantic Consistency*, directly supervise prompt-to-region alignment.
- HPI achieves state-of-the-art performance on synthetic-to-real (70.62 mIoU) and real-to-real (72.74 mIoU) DGSS benchmarks, outperforming strong single-foundation and multi-model baselines on average.

2 Related Work

Domain generalized semantic segmentation (DGSS) DGSS seeks to train on source domains and generalize to unseen targets without accessing target data [39]. Early work reduces the domain gap at the surface level through style

randomization [1, 22, 36, 63, 64], frequency perturbation [6, 10, 31, 47, 55], and texture diversification [25, 28, 29]. A complementary direction explores semantic invariance against style shifts directly in the feature space through domain-invariant feature learning [37, 51], hierarchical structural grouping [18, 27], and uncertainty-aware feature modeling [11, 30]. However, these methods typically rely on conventional backbones with limited pre-training scope, constraining the breadth of visual knowledge available for generalization.

The availability of pretrained large models [34, 38] has shifted the paradigm toward transferring their knowledge into DGSS, such as through parameter-efficient adapters [7, 21, 24, 45, 48, 58, 62] or multi-model fusion [5, 12, 60]. Among these, methods that inject rich semantics into visual features through VLMs [5, 23,35,40,60] have demonstrated a clear advantage, as the semantic priors encoded in VLMs are inherently less sensitive to low-level appearance variation and thus provide more domain-invariant guidance for dense prediction. This motivates us to further exploit VLM semantics by designing spatially structured prompts and an adaptive injection mechanism that together push the boundary of semantic-guided DGSS.

Prompt learning for domain generalization Building on VLMs, prompt learning has become a widely adopted strategy for injecting domain-invariant semantic priors into visual representations [65, 66]. To enhance cross-domain robustness, some methods utilize prompts to diversify prompt content to simulate unseen domain shifts [2, 15, 46, 49, 52], disentangle prompt representations based on domain properties [3, 8, 14, 23, 26, 50, 59], or detect open-set categories [9, 43]. However, all these methods encode semantics at the class level and ignore spatially aligned and adaptive injection.

When extending prompt learning to semantic segmentation, a standard practice is to integrate textual semantics into the decoder stage, such as semantic prompt queries [33] or text-to-pixel attention [35]. However, these methods leave the most important visual encoder completely unguided by semantic priors. Therefore, recent approaches leverage prompt learning for the encoder, actively modulating intermediate visual features through pixel-text score maps [40], non-linear text-to-vision projections [4], or multi-scale feature alignments [53, 57]. Despite this progression, current frameworks exhibit two structural limitations: they rely on holistic class-level prompts that overlook intra-class spatial diversity, and they apply uniform semantic injection that risks negative transfer in background or occluded regions. Our work addresses both gaps through region-level geometric prompts and adaptive confidence-gated injection.

3 Method

As illustrated in Fig. 2, we first generate SHP using an LLM and encode them with the VLM text encoder E_T to obtain structural prompt features f_{SHP}. Given an input image, the frozen VLM image encoder E_I^{vlm} extracts visual features $\mathbf{P}^{\text{vlm}} \in \mathbb{R}^{B\times N\times C}$ as the primary backbone, where B, N, C denote the batch size, number of patches, and channel dimension. In parallel, we feed

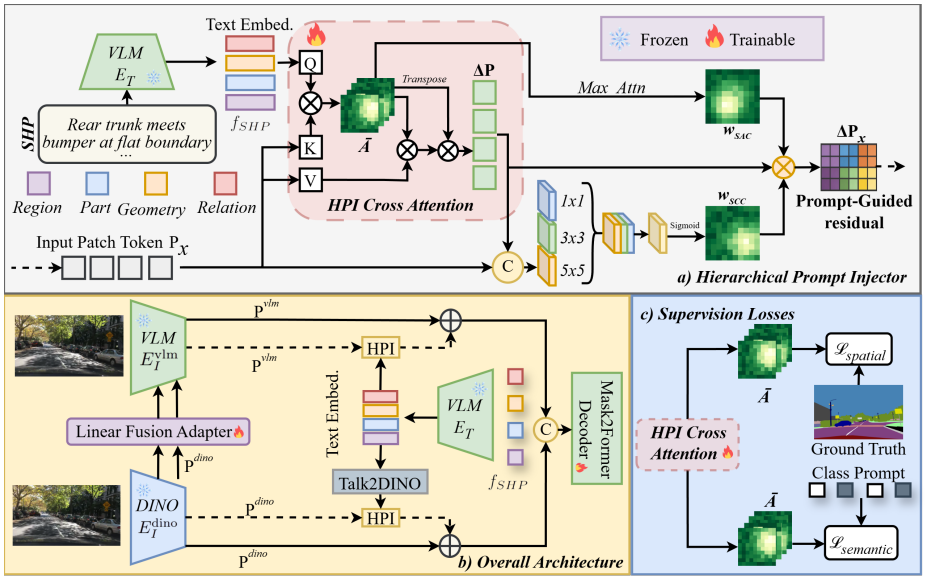

Fig. 2: (a) Hierarchical Prompt Injector. Region-level prompts serve as queries in cross-attention, and the resulting residuals are modulated by SAC and SCC before injection. **(b)** Overall architecture. HPI augments the VLM backbone with SHP-guided residuals, while DINOv2 provides spatial features that are fused into the VLM stream for dense prediction. **(c)** Supervision losses. $\mathcal{L}_{\text{spatial}}$ aligns each prompt's cross-attention map $\bar{\mathbf{A}}^x$ with ground-truth segmentation masks, supervising prompt-to-region correspondences. $\mathcal{L}_{\text{semantic}}$ supervises class-level logits $\hat{\mathbf{s}}$ against binary class-presence labels.

the same image I into a frozen DINOv2 encoder E_I^{dino} [34], producing features $\mathbf{P}^{\text{dino}} \in \mathbb{R}^{B\times N\times C}$ that provide geometrically precise cues for spatial grounding in E_I^{vlm}. $\mathbf{P}^{\text{dino}}$ is fused into the VLM stream through linear adapters [21], allowing fine-grained spatial structure to continuously refine VLM representations. At deep layers of E_I^{vlm} and E_I^{dino}, HPI applies cross-attention between f_{SHP} and $\mathbf{P}^x$, where $x \in \{\text{vlm}, \text{dino}\}$, to produce prompt-guided residuals $\Delta\mathbf{P}^x$ that carry region-level geometric cues grounded onto each spatial position. Two patch-wise confidence maps, SAC and SCC, selectively reweight the prompt-guided residual, ensuring that the updates are semantically relevant and spatially coherent. Finally, we introduce a *Spatial Localization Loss* $\mathcal{L}_{\text{spatial}}$ to encourage each region-level prompt to attend to its ground-truth object region, and a *Semantic Consistency Loss* $\mathcal{L}_{\text{semantic}}$ to suppress activations of classes absent from the image. Multi-scale features from both E_I^{vlm} and E_I^{dino} are concatenated and decoded by a Mask2Former head [13] to predict segmentation results.

3.1 Spatial Hierarchical Prompts

Class-level holistic prompts provide a single textual anchor per category. However, in real driving scenes, viewpoint changes and inter-object occlusion rou-

tinely hide large portions of an object. A single global description therefore cannot reliably match the visible content. Our key idea is to decompose each class into a small set of region-level prompts that capture complementary geometric facets, so that at least one prompt remains aligned with the visible content under arbitrary viewing conditions. We design textual prompts that (1) cover complementary spatial regions of the object and (2) encode appearance-invariant geometric cues within each region. To achieve this, we define R regional prompts per class c_i: $\mathcal{D}_i = \{d_i^{(r)}\}_{r=1}^{R}$, where each $d_i^{(r)}$ describes a spatial region chosen to form a minimal partition of the object (e.g. front/side/rear for a vehicle). As illustrated in Fig. 1, we further design each prompt $d_i^{(r)}$ using four complementary elements that define both the geometrical location and structural information of an object region: [Region Cue] + [Part] + [Geometry] + [Boundary Relation].

The region cue anchors the viewing angle so that HPI can select the appropriate prompt in each $\mathcal{D}_i$ at inference time. Part and geometry together identify a discriminative, appearance-invariant local structure. The boundary relation grounds the region relative to its neighbors, providing contextual cues that aid segmentation at object borders. For example, a *rider* prompt—"Head forms round helmet above handlebar boundary"—captures region-specific geometric cues that persist across domains. Each class is partitioned into R such complementary descriptions, ensuring that at least one remains aligned with the visible content under arbitrary viewpoint changes. We set R=3 by default and validate this choice in Tab. 3. Each prompt is wrapped in a unified template:

$$\text{prompt}(c_i, d_i^{(r)}) = \text{"a photo of a } c_i \text{ which } d_i^{(r)}\text{"}. \tag{1}$$

We generate all $K{=}C_{\text{cls}}{\times}R$ prompts offline using an LLM with structured instructions specifying the four components for each class–region combination (see *Supplementary* for the full generation protocol and prompt list). The resulting structural prompt features $f_{SHP} \in \mathbb{R}^{K\times D}$ are obtained by encoding each prompt once with the frozen VLM text encoder E_T, where D is the text embedding dimension.

3.2 Hierarchical Prompt Injector Framework

Given f_{SHP} from Sec. 3.1, a key challenge is to turn these region-level descriptions into *hierarchical, spatially localized* guidance for dense features. With only f_{SHP}, directly reusing prior prompt-injection paradigms makes it difficult to exploit the region hierarchy. These previous methods typically treat text as a single global condition or as decoder-only queries, where the prompt signal is either absent from the feature extractor or applied without patch-specific selectivity. To address this, HPI employs cross-attention between prompt embeddings and patch tokens to generate prompt-conditioned residual updates $\Delta\mathbf{P}^x$. As shown in Fig. 2(b), this spatially selective routing enables different region-level prompts to guide their corresponding visual patches effectively.

Hierarchical Prompt Injection We first prepare specific prompt embeddings for E_I^{vlm} and E_I^{dino} before injecting them into both encoders. Specifically, we

zero-pad $f_{SHP} \in \mathbb{R}^{K\times D}$ along the channel dimension to obtain $f_{SHP}^{\text{vlm}} \in \mathbb{R}^{K\times C}$, since E_I^{vlm} and E_T are jointly trained with aligned representations. However, no such text-visual alignment exists in DINOv2, as it is trained with purely visual self-supervision. We thus project f_{SHP} into the DINOv2 feature space via Talk2DINO [4], a pretrained text-to-vision projection, yielding $f_{SHP}^{\text{dino}} \in \mathbb{R}^{K\times C}$. We generate f_{SHP}^{vlm} and f_{SHP}^{dino} once before training and keep them fixed, incurring minimal cost.

For branch $x \in \{\text{vlm}, \text{dino}\}$, let $\mathbf{P}^x$ denote the patch-token output at the designated injection layer. We compute cross-attention between f_{SHP}^x and $\mathbf{P}^x$:

$$\mathbf{A}^x = \text{CrossAttn}(f_{SHP}^x,\ \mathbf{P}^x) \in \mathbb{R}^{B\times K\times N}, \tag{2}$$

Finally, to generate the prompt-guided residual $\Delta\mathbf{P}^x$, we compute the dot product between $\mathbf{A}^x$ and its transpose, which is multiplied by the value $\mathbf{V}^x$ in HPI cross attention:

$$\Delta\mathbf{P}^x = (\mathbf{A}^x)^\top\, \mathbf{A}^x\, \mathbf{V}^x\ \in\ \mathbb{R}^{B\times N\times C}. \tag{3}$$

In Eq. 3, $(\mathbf{A}^x)^\top\mathbf{A}^x$ forms a prompt-induced patch routing matrix, where patches are strongly connected when they are co-activated by the same region-level prompts; multiplying this matrix with $\mathbf{V}^x$ then aggregates visual values among prompt-related patches. This differs from standard cross-attention by routing visual-space information rather than directly injecting prompt-valued residuals.

Adaptive Injection Although cross-attention assigns data-dependent weights, the softmax normalization allocates non-zero mass to all keys, so the prompt-guided residual $\Delta\mathbf{P}^x$ can still leak into patches where the described region is absent or heavily occluded, introducing spurious feature updates. Moreover, cross-attention lacks an explicit 2D continuity prior, providing no guarantee that highly activated patches form spatially coherent regions. We therefore compute two complementary patch-wise confidence weights, namely Semantic Alignment Confidence (SAC) $\mathbf{w}_{\text{SAC}} \in [0,1]^{B\times N\times 1}$ and Spatial Coherence Confidence (SCC) $\mathbf{w}_{\text{SCC}} \in [0,1]^{B\times N\times 1}$. These weights selectively re-weight $\Delta\mathbf{P}^x$, suppressing misaligned patches while promoting spatially coherent refinement. Specifically, from the cross-attention in Eq. (2), we extract the head-averaged attention weights $\bar{\mathbf{A}}^x \in \mathbb{R}^{B\times N\times K}$, which capture patch–prompt affinities.

$$\mathbf{w}_{\text{SAC},b,n} = \max_{k\in[K]} \bar{\mathbf{A}}^x_{b,n,k}. \tag{4}$$

Therefore, patches where no prompt is semantically relevant receive low $\mathbf{w}_{\text{SAC}}$, suppressing erroneous injection.

However, SAC evaluates each patch independently, so a noisy patch may receive a high attention score when two objects share similar parts (e.g. wheels on trucks and cars), leading to undesirably high $\mathbf{w}_{\text{SAC}}$ for misaligned patches. SCC addresses this by incorporating multi-scale 2D spatial context, allowing it to verify whether a high-SAC patch is supported by its local neighborhood rather than being an isolated spurious activation. Specifically, we concatenate $[\mathbf{P}^x,\ \Delta\mathbf{P}^x]$ and project it to a lower-dimensional bottleneck, then reshape the result to a 2D spatial grid and apply parallel convolutions at kernel sizes 1×1, 3×3,

and 5×5 followed by a 1×1 fusion convolution and sigmoid function, producing $\mathbf{w}_{\mathrm{SCC}} \in [0,1]^{B\times N\times 1}$. The three kernel sizes capture spatial context at different receptive fields, accommodating the large variation in object scale across driving-scene categories.

Since $\mathbf{w}_{\mathrm{SAC}}$ derives from attention statistics and $\mathbf{w}_{\mathrm{SCC}}$ derives from spatial convolutions, the two weights carry complementary information but may differ in magnitude. We introduce a learnable scalar $\beta = \sigma(\beta_0) \in (0,1)$, where σ is the sigmoid function and β_0 is initialized to 0 so that β starts at 0.5, to balance their contributions. The final prompt-enhanced features are

$$\mathbf{P'}^{x} = \mathbf{P}^{x} + \left[(1-\beta)\,\mathbf{w}_{\mathrm{SAC}} + \beta\,\mathbf{w}_{\mathrm{SCC}}\right] \odot \Delta\mathbf{P}^{x}. \tag{5}$$

3.3 Supervision for Prompt Alignment

The cross-attention in Eq. (2) produces prompt-to-patch correspondences, yet these correspondences may not accurately localize each region-level description without additional guidance, because VLMs trained with image-level contrastive objectives lack explicit patch-level alignment [5,35]. We therefore introduce two auxiliary losses that supervise these correspondences at complementary granularities: $\mathcal{L}_{\mathrm{spatial}}$ aligns each prompt's attention map with the ground-truth segmentation mask at the pixel level, while $\mathcal{L}_{\mathrm{semantic}}$ provides complementary image-level supervision that suppresses activations of classes absent from the image.

Spatial Localization Loss To ensure that HPI injects region-level cues only where the corresponding object actually appears, we use the ground truth label y_{seg} as target to constrain the attention weight $\bar{\mathbf{A}}^{x} \in \mathbb{R}^{B\times N\times K}$ in HPI. This is because y_{seg} encodes exact object boundaries at each spatial position, while $\bar{\mathbf{A}}^{x}$ determines how each prompt routes the residual $\Delta\mathbf{P}^{x}$ to patches. Aligning the two forces the HPI to concentrate updates on the correct object regions. Since $\bar{\mathbf{A}}^{x}$ has K prompt-level entries per patch whereas y_{seg} is defined over C_{cls} classes, we first aggregate prompt-level affinities into a patch-wise class distribution $\boldsymbol{\pi}^{x} \in \mathbb{R}^{B\times C_{\mathrm{cls}}\times H\times W}$ that can be directly supervised by y_{seg}. Let $\mathcal{K}(c)$ denote the R prompt indices for class c. We compute $\boldsymbol{\pi}^{x}$ by aggregating the regional attentions per class via log-sum-exp and normalizing across classes with a softmax:

$$\boldsymbol{\pi}^{x}_{c}(n) = \frac{\exp(\mathrm{LSE}_{r\in\mathcal{K}(c)}(\log \bar{\mathbf{A}}^{x}_{n,r}))}{\sum_{c'}\exp(\mathrm{LSE}_{r\in\mathcal{K}(c')}(\log \bar{\mathbf{A}}^{x}_{n,r}))} \tag{6}$$

where LSE denotes $\log\sum\exp$. Finally, we upsample $\boldsymbol{\pi}^{x}$ to the same shape as the segmentation label y_{seg} and compute the negative log likelihood loss between them

$$\mathcal{L}_{\mathrm{spatial}} = \sum_{x\in\{\mathrm{vlm},\,\mathrm{dino}\}} \mathrm{NLL}((\boldsymbol{\pi}^{x})_{\uparrow},\ y_{\mathrm{seg}}). \tag{7}$$

Semantic Consistency Loss $\mathcal{L}_{\mathrm{spatial}}$ steers each prompt to its correct region, yet the softmax in Eq. (6) can only implicitly discourage activations for absent

classes and cannot drive them to zero. We therefore use a binary class-presence vector $\mathbf{y}_{\text{cls}} \in \{0,1\}^{B\times C_{\text{cls}}}$, derived from y_{seg}, to constrain the attention weight $\bar{\mathbf{A}}^x$ in HPI to explicitly suppress injection for absent classes. This is because $\mathbf{y}_{\text{cls}}$ directly encodes whether each class appears in the image, while $\bar{\mathbf{A}}^x$ reveals what each prompt actually attends to. Comparing the two exposes prompts that activate on absent classes. Since $\bar{\mathbf{A}}^x$ operates at prompt granularity whereas $\mathbf{y}_{\text{cls}}$ is defined over C_{cls} classes, we aggregate prompt-level evidence into class-level logits $\hat{\mathbf{s}} \in \mathbb{R}^{B\times C_{\text{cls}}}$ that indicate how likely each class is present and can be directly supervised by $\mathbf{y}_{\text{cls}}$. We first compute a per-prompt feature $\mathbf{v}^x_{b,k}$ by aggregating patch features weighted by the attention distribution:

$$\mathbf{v}^x_{b,k} = \sum_{n=1}^{N} \bar{\mathbf{A}}^x_{b,n,k}\, \mathbf{P}^x_{b,n} \;\in\; \mathbb{R}^C. \tag{8}$$

We then measure the cosine similarity between $\mathbf{v}^x_{b,k}$ and its prompt embedding $f^x_{SHP,k}$, scale by a learnable temperature, and mean-pool the R regional scores per class to obtain $\hat{\mathbf{s}}$. We compute binary cross-entropy loss between $\hat{\mathbf{s}}$ and $\mathbf{y}_{\text{cls}}$:

$$\mathcal{L}_{\text{semantic}} = \sum_{x\in\{\text{vlm},\,\text{dino}\}} \text{BCE}(\hat{\mathbf{s}}^x,\; \mathbf{y}_{\text{cls}}). \tag{9}$$

Full objective The final training objective combines the Mask2Former segmentation loss with the two prompt-level losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{spatial}}\,\mathcal{L}_{\text{spatial}} + \lambda_{\text{semantic}}\,\mathcal{L}_{\text{semantic}}, \tag{10}$$

where $\mathcal{L}_{\text{seg}}$ follows the standard Mask2Former formulation [13].

4 Experiments

4.1 Experimental Settings

Datasets We conduct experiments on five driving-scene semantic segmentation datasets that share 19 common semantic classes, and use the abbreviations in parentheses to denote them throughout the paper. GTA5 [41] (G) contains 24,966 training images at a resolution of 1914×1052. SYNTHIA [42] (S) provides 9,400 training images with a resolution of 1280×760. Cityscapes [17] (C) includes 2,975 training and 500 validation images at a resolution of 2048×1024. BDD100K [56] (B) consists of 1,000 validation images at a resolution of 1280×720, and Mapillary Vistas [32] (M) comprises 2,000 validation images with varying resolutions.
Evaluation Following existing DGSS methods [16, 37, 48, 60], we use A→B to indicate training on dataset A and evaluating on dataset B. We consider three standard settings: (1) G→{C, B, M}; (2) S→{C, B, M}; and (3) C→{B, M}.
Implementation details Following [35, 48, 60], we freeze the pretrained foundation models and tune only the HPI modules, linear fusion adapters, and the Mask2Former decoder. Inputs are resized to 512×512. We train with batch

Table 1: Performance comparison (mIoU) under the synthetic-to-real (G→{C, B, M}) and real-to-real (C→{B, M}) settings. † denotes methods that additionally employ a frozen DINOv2-L encoder. Best and second-best in **bold** and underlined.

Methods	Backbone	Synthetic-to-real				Real-to-real		
		G→C	G→B	G→M	Avg.	C→B	C→M	Avg.
SAN-SAW [37]	RN101	45.33	41.18	40.77	42.43	54.73	61.27	58.00
WildNet [29]	RN101	45.79	41.73	47.08	44.87	47.01	50.94	48.98
SHADE [63]	RN101	46.66	43.66	45.50	45.27	50.95	60.67	55.81
TLDR [28]	RN101	47.58	44.88	48.80	47.09	–	–	–
FAMix [19]	RN101	49.47	46.40	51.97	49.28	54.07	58.72	56.40
Rein [48]	DINOv2-L	66.40	60.40	66.10	64.30	65.00	72.30	68.65
SET [55]	DINOv2-L	68.06	61.64	67.68	65.79	65.07	75.67	70.37
FADA [7]	DINOv2-L	68.23	61.94	68.09	66.09	65.12	75.86	70.49
FisherTune [62]	DINOv2-L	68.20	**63.30**	68.70	66.73	–	–	–
DepthForge [12]	DINOv2-L	69.04	62.82	69.22	67.02	66.19	75.93	71.06
SoMA [58]	DINOv2-L	71.82	61.31	71.67	68.27	**67.02**	76.45	71.74
CLOUDS† [5]	CLIP-L	60.20	57.40	67.00	61.50	–	–	–
MFuser† [60]	CLIP-L	71.24	61.08	71.14	67.82	65.58	78.10	71.84
HPI (Ours)†	**CLIP-L**	**74.05**	63.14	**74.67**	**70.62**	66.01	**79.44**	72.73
TQDM [35]	EVA02-L	68.88	59.18	70.10	66.05	64.72	76.15	70.44
DPMFormer [23]	EVA02-L	70.08	60.48	70.66	67.07	64.20	76.67	70.44
MFuser† [60]	EVA02-L	70.19	63.13	71.28	68.20	65.81	77.93	71.87
HPI (Ours)†	**EVA02-L**	72.86	62.52	72.80	69.39	66.23	79.25	**72.74**

size 2 and learning rate 1e−4 for 12k iterations. AdamW optimizer is employed with a linear warm-up over 1.5k iterations. We set the spatial localization weight λ_{spatial}=0.1 and the semantic consistency weight $\lambda_{\text{semantic}}$=0.05. All experiments are conducted on one NVIDIA RTX 3090 GPU.

Comparison methods We compare HPI against representative DGSS methods under all three evaluation protocols. Specifically, we include recent single-foundation-model approaches, including Rein [48], SET [55], FADA [7], TQDM [35], DPMFormer [23], FisherTune [62], and SoMA [58]. We further consider multi-foundation-model frameworks that integrate multiple pretrained models, such as CLOUDS [5], MFuser [60], and DepthForge [12]. For completeness, we also report earlier CNN-based methods, including SAN-SAW [37], WildNet [29], SHADE [63], TLDR [28], and FAMix [19].

4.2 Main Results

G→{C, B, M} Tab. 1 compares HPI with representative DGSS methods under the synthetic-to-real setting. HPI achieves the best average mIoU and attains the highest scores on Cityscapes and Mapillary, while remaining competitive on BDD100K. With the CLIP-L backbone, we achieve an average mIoU of 70.62, improving over MFuser by 2.80 and over the strongest single-model method SoMA by 2.35 on average. The gain is particularly pronounced on Mapillary, where we improve over SoMA by 3.00 mIoU. We attribute this to the region-level spatial decomposition in SHP, which provides geometric anchors that remain discriminative under the large viewpoint and layout variations present in Mapillary, whereas class-level holistic descriptions used in prior methods lack the spatial

Table 2: Performance comparison (mIoU) under the synthetic-to-real setting (S→{C, B, M}). † denotes methods that additionally employ a frozen DINOv2-L encoder. Best and second-best in **bold** and underlined.

Methods	Backbone	Synthetic-to-real			
		S→C	S→B	S→M	Avg.
SAN-SAW [37]	RN101	40.87	35.98	37.26	38.04
TLDR [28]	RN101	42.60	35.46	37.46	38.51
IBAFormer [44]	MiT-B5	50.92	44.66	50.58	48.72
Rein [48]	DINOv2-L	48.59	44.42	48.64	47.22
SET [55]	DINOv2-L	49.65	45.45	49.45	48.18
FADA [7]	EVA02-L	50.04	45.83	49.86	48.57
MFuser† [60]	EVA02-L	54.17	46.67	53.22	51.35
CLOUDS† [5]	CLIP-L	44.97	39.58	46.99	43.85
MFuser† [60]	CLIP-L	53.16	45.59	52.27	50.34
HPI (Ours)†	**CLIP-L**	**55.00**	**49.19**	**53.98**	**52.72**

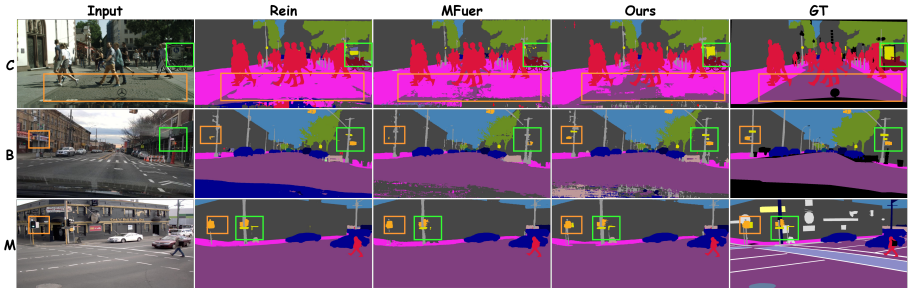

Fig. 3: Qualitative comparison on G→{C, B, M}. From left to right — input image, Rein [48], MFuser [60], HPI (Ours), and ground truth. Highlighted boxes indicate regions with partial occlusion or fine-grained structures where prior methods produce erroneous predictions.

grounding needed to resolve such diversity. The adaptive SAC–SCC gating further ensures that only semantically relevant prompts are injected at spatially coherent locations, preventing noisy activations from degrading predictions in cluttered scenes.

C→{B, M} With the EVA02-L backbone, HPI achieves 72.74 average mIoU, surpassing SoMA by 1.00 and MFuser by 0.87. With CLIP-L, HPI attains the highest single-target score of 79.44 mIoU on Mapillary, confirming that the gains transfer across backbone choices.

S→{C, B, M} We additionally train on SYNTHIA to verify robustness across different synthetic source domains. As reported in Tab. 2, HPI improves over MFuser (CLIP-L) by 2.38 mIoU on average, with a particularly large gain of 3.60 mIoU on BDD100K. We attribute this to SHP providing partially matchable region-level geometric anchors under SYNTHIA-to-real appearance mismatch, while the SAC–SCC gating confines injection to semantically relevant and spatially coherent regions, reducing background leakage in cluttered scenes.

Table 3: Spatial hierarchical prompt ablation study (mIoU) with DINOv2-L+CLIP-L.

Configuration	G→C	G→B	G→M	Avg.
Number of regional prompts (R)				
R=2	72.68	62.26	73.49	69.48
R=4	73.01	62.58	73.48	69.69
Geometrical Cues				
w/o Region Cue	72.33	62.24	73.26	69.28
w/o Part	72.49	62.40	73.30	69.40
w/o Geometry	72.76	62.40	73.17	69.44
w/o Boundary Relation	72.96	62.50	73.77	69.74
Ours (R=3, full prompt)	**74.05**	**63.14**	**74.67**	**70.62**

Table 4: Adaptive injection ablation study (mIoU) with DINOv2-L+CLIP-L.

Configuration	G→C	G→B	G→M	Avg.
w/o SAC	73.83	61.61	73.76	69.73
w/o SCC	73.17	62.63	74.03	69.94
Ours	**74.05**	**63.14**	**74.67**	**70.62**

Qualitative analysis Fig. 3 presents qualitative comparisons on G→{C, B, M}. We highlight regions where prior methods struggle, particularly under partial occlusion. HPI consistently produces more accurate and spatially coherent masks for partially occluded objects such as vehicles and pedestrians.

4.3 In-Depth Analysis

Number of regional prompts To determine the optimal granularity of spatial decomposition, we vary R, the number of regional prompts per class, from 2 to 4. As shown in Tab. 3, performance peaks at R=3 with 70.62 average mIoU and drops at both R=2 with 69.48 and R=4 with 69.69. This indicates that an insufficient number of regions leaves substantial portions of an object uncovered under viewpoint variation, whereas an excessive number introduces overlapping descriptions whose cross-attention responses compete, diluting selectivity. We use R=3 throughout all other experiments.

SHP construction Each regional prompt follows the structured formulation described in Sec. 3.1. We ablate each component by removing it from all prompts. Removing Region Cue incurs the largest drop of 1.34 mIoU, as it serves as the viewpoint anchor that enables HPI to select the appropriate regional description at inference time. Part and Geometry cause drops of 1.22 and 1.18 mIoU, respectively, together capturing discriminative, appearance-invariant local structures that persist across domains. Boundary Relation incurs the smallest drop of 0.88 mIoU, consistent with its auxiliary role of providing contextual grounding at object borders.

Adaptive injection We ablate the proposed adaptive gating by removing SAC or SCC from HPI (Tab. 4). Removing SAC causes a drop of 0.89 mIoU, as without semantic filtering the injector activates prompts indiscriminately, allowing irrelevant region descriptions to corrupt patch features under domain

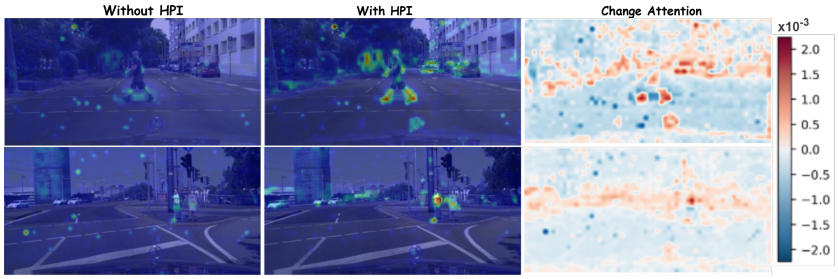

Fig. 4: Cross-attention maps before (left) and after (middle) HPI injection, with the difference map (right). HPI concentrates attention on semantically meaningful object parts while suppressing background activations.

Table 5: HPI component ablation study (mIoU). We additionally report CLIP/EVA-only baselines from Rein [48] for reference under the same G→{C,B,M} setting.

Configuration	CLIP-L				EVA02-L			
	G→C	G→B	G→M	Avg.	G→C	G→B	G→M	Avg.
w/o HPI	71.57	61.29	72.98	68.61	71.63	**63.48**	72.08	69.06
w/o DINOv2	61.58	53.60	65.81	60.33	69.20	60.67	70.01	66.63
Freeze [48]	53.70	48.70	55.00	52.40	56.50	53.60	58.60	56.20
Rein [48]	57.10	54.70	60.50	57.40	65.30	60.50	64.90	63.60
SET [55]	58.20	55.30	61.40	58.30	66.40	61.80	65.60	64.60
Ours	**74.05**	**63.14**	**74.67**	**70.62**	**72.86**	62.52	**72.80**	**69.39**

shift. Moreover, removing SCC leads to a drop of 0.68 mIoU, as without spatial coherence enforcement the injection produces isolated noisy activations that fragment the injection map rather than forming contiguous object regions. The two mechanisms serve complementary roles, where SAC determines whether a prompt should activate at each patch and SCC refines the spatial extent of active prompts to ensure coherent regions rather than scattered responses.

Fig. 4 visualizes the cross-attention maps before and after HPI injection. Without HPI, the attention is diffuse and spread across background regions. After injection, the attention concentrates on semantically meaningful object parts—vehicle bodies, pedestrian silhouettes, and structural boundaries. The difference map confirms that HPI systematically strengthens foreground responses while suppressing background activations, consistent with the design that region-level geometric anchors guide the model to attend to structurally informative regions under domain shift.

Foundation model complementarity To isolate the contribution of DINOv2, we disable the DINOv2 branch and apply HPI to the VLM encoder alone. As shown in Tab. 5, with EVA02-L only, HPI achieves 66.63 average mIoU, already exceeding representative single-foundation baselines such as SET and Rein listed in the same table. With CLIP-L only, HPI reaches 60.33 average mIoU, indicating that region-level prompts and adaptive injection remain effective even without a

Table 6: Injection layer ablation study (mIoU) with DINOv2-L+CLIP-L.

Layer	G→C	G→B	G→M	Avg.	Layer	G→C	G→B	G→M	Avg.
5	72.29	62.25	73.43	69.32	17	72.79	62.72	73.75	69.75
9	72.66	62.43	73.49	69.53	21	73.15	62.98	73.59	69.91
13	72.59	62.78	73.36	69.58	Ours (23)	**74.05**	**63.14**	**74.67**	**70.62**

secondary encoder. Adding DINOv2 further improves the average by 2.76 mIoU for EVA02-L and 10.29 mIoU for CLIP-L, highlighting the complementary role of spatially grounded features. The larger gain for CLIP-L likely reflects its weaker native spatial encoding relative to EVA02-L.

Injection layer Tab. 6 sweeps the HPI insertion layer from 5 to 23. Performance increases monotonically with depth, peaking at layer 23, as shallower layers encode low-level texture features that are domain-sensitive and poorly aligned with the semantic content of region-level prompts [54]. Deeper layers instead provide mature class-discriminative representations where geometric anchors can take effect most precisely.

Table 7: Supervision loss ablation study (mIoU) with DINOv2-L+CLIP-L.

Configuration	G→C	G→B	G→M	Avg.
w/o $\mathcal{L}_{\text{spatial}}$ + $\mathcal{L}_{\text{semantic}}$	72.51	62.69	73.01	69.40
w/o $\mathcal{L}_{\text{spatial}}$	72.99	62.07	73.56	69.54
w/o $\mathcal{L}_{\text{semantic}}$	73.10	62.87	73.76	69.91
Ours	**74.05**	**63.14**	**74.67**	**70.62**

Supervision losses As shown in Tab. 7, removing both losses reduces the average mIoU to 69.40, a drop of 1.22 mIoU. Removing $\mathcal{L}_{\text{spatial}}$ leads to a 1.08 mIoU drop, demonstrating that explicit spatial supervision is essential for guiding cross-attention toward correct object regions. Similarly, excluding $\mathcal{L}_{\text{semantic}}$ results in a 0.71 mIoU decline, which suggests its critical role in regularizing prompt activations to prevent interference from categories absent in the scene.

5 Conclusion

In this paper, we propose spatial hierarchical prompts that describe each class with a small set of region-level phrases capturing stable geometric cues, part–boundary relations, and visibility patterns, serving as geometric anchors under domain shift. Building on this design, we introduce HPI, which uses these regional prompts to robustly steer frozen vision–language and vision foundation model features toward spatially consistent, domain-robust predictions. Together with adaptive injection that controls *whether* and *how much* each prompt takes effect, HPI stabilizes vision–language alignment across domains, making frozen foundation models effective for dense prediction under distribution shift.

References

1. Ahn, W.J., Yang, G.Y., Choi, H.D., Lim, M.T.: Style blind domain generalized semantic segmentation via covariance alignment and semantic consistence contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3616–3626 (2024)
2. Bai, S., Zhang, Y., Zhou, W., Luan, Z., Chen, B.: Soft prompt generation for domain generalization. In: European Conference on Computer Vision. pp. 434–450. Springer (2024)
3. Bai, S., Zhang, J., Guo, S., Li, S., Guo, J., Hou, J., Han, T., Lu, X.: Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27284–27293 (2024)
4. Barsellotti, L., Bianchi, L., Messina, N., Carrara, F., Cornia, M., Baraldi, L., Falchi, F., Cucchiara, R.: Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22025–22035 (2025)
5. Benigmim, Y., Roy, S., Essid, S., Kalogeiton, V., Lathuilière, S.: Collaborating foundation models for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3108–3119 (2024)
6. Bi, Q., Yi, J., Huang, H., Zheng, H., Zhan, H., Huang, Y., Li, Y., Wu, X., Zheng, Y.: Nightadapter: Learning a frequency adapter for generalizable night-time scene segmentation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23838–23849 (2025)
7. Bi, Q., Yi, J., Zheng, H., Zhan, H., Huang, Y., Ji, W., Li, Y., Zheng, Y.: Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 94047–94072. Curran Associates, Inc. (2024)
8. Bose, S., Jha, A., Fini, E., Singha, M., Ricci, E., Banerjee, B.: Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5542–5552 (2024)
9. C, M.H.N., Gupta, D., Singha, M., Rongali, S.B., Jha, A., Khan, M.H., Banerjee, B.: Osloprompt: Bridging low-supervision challenges and open-set domain generalization in clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10110–10120 (June 2025)
10. Chattopadhyay, P., Sarangmath, K., Vijaykumar, V., Hoffman, J.: Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19288–19300 (2023)
11. Chen, I.H., Chang, H.E., Chen, W.T., Hwang, J.N., Kuo, S.Y.: Exploring probabilistic modeling beyond domain generalization for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21755–21765 (2025)
12. Chen, S., Han, T., Zhang, C., Luo, X., Wu, M., Cai, G., Su, J.: Stronger, steadier & superior: Geometric consistency in depth vfm forges domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8285–8295 (2025)

13. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
14. Cheng, D., Xu, Z., Jiang, X., Wang, N., Li, D., Gao, X.: Disentangled prompt representation for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23595–23604 (2024)
15. Cho, J., Nam, G., Kim, S., Yang, H., Kwak, S.: Promptstyler: Prompt-driven style generation for source-free domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15702–15712 (2023)
16. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11580–11590 (2021)
17. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
18. Ding, J., Xue, N., Xia, G.S., Schiele, B., Dai, D.: Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15413–15423 (2023)
19. Fahes, M., Vu, T.H., Bursuc, A., Pérez, P., De Charette, R.: A simple recipe for language-guided domain generalized segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23428–23437 (2024)
20. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International conference on learning representations (2019)
21. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799 (2019)
22. Huang, W., Chen, C., Li, Y., Li, J., Li, C., Song, F., Yan, Y., Xiong, Z.: Style projected clustering for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3061–3071 (2023)
23. Jeon, S., Hong, K., Byun, H.: Exploiting domain properties in language-driven domain generalization for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20791–20801 (2025)
24. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727 (2022)
25. Jia, Y., Hoyer, L., Huang, S., Wang, T., Van Gool, L., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: European Conference on Computer Vision. pp. 91–109 (2024)
26. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)

27. Kim, J., Lee, J., Park, J., Min, D., Sohn, K.: Pin the memory: Learning to generalize semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4350–4360 (2022)
28. Kim, S., Kim, D.h., Kim, H.: Texture learning domain randomization for domain generalized segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 677–687 (2023)
29. Lee, S., Seong, H., Lee, S., Kim, E.: Wildnet: Learning domain generalized semantic segmentation from the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9936–9946 (2022)
30. Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., Duan, L.Y.: Uncertainty modeling for out-of-distribution generalization. In: International Conference on Learning Representations (2022)
31. Lin, S., Zhang, Z., Huang, Z., Lu, Y., Lan, C., Chu, P., You, Q., Wang, J., Liu, Z., Parulkar, A., et al.: Deep frequency filtering for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11797–11807 (2023)
32. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision. pp. 4990–4999 (2017)
33. Niu, H., Xie, L., Lin, J., Zhang, S.: Exploring semantic consistency and style diversity for domain generalized semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6245–6253 (2025)
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
35. Pak, B., Woo, B., Kim, S., Kim, D.h., Kim, H.: Textual query-driven mask transformer for domain generalized segmentation. In: European Conference on Computer Vision. pp. 37–54 (2024)
36. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the european conference on computer vision (ECCV). pp. 464–479 (2018)
37. Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W.: Semantic-aware domain generalized segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2594–2605 (2022)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
39. Rafi, T.H., Mahjabin, R., Ghosh, E., Ko, Y.W., Lee, J.G.: Domain generalization for semantic segmentation: a survey. Artificial Intelligence Review **57**(9), 247 (2024)
40. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18082–18091 (2022)
41. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European Conference on Computer Vision. pp. 102–118 (2016)

42. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
43. Singha, M., Jha, A., Bose, S., Nair, A., Abdar, M., Banerjee, B.: Unknown prompt the only lacuna: Unveiling clip's potential for open domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13309–13319 (2024)
44. Sun, Q., Chen, H., Zheng, M., Wu, Z., Felsberg, M., Tang, Y.: Ibaformer: Intra-batch attention transformer for domain generalized semantic segmentation. arXiv preprint arXiv:2309.06282 (2023)
45. Tang, P., Zhang, X., Yang, C., Yuan, H., Sun, J., Shan, D., Yang, Z.J.: Unleashing the power of visual foundation models for generalizable semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 20823–20831 (2025)
46. Tang, Y., Wan, Y., Qi, L., Geng, X.: Dpstyler: dynamic promptstyler for source-free domain generalization. IEEE Transactions on Multimedia **27**, 120–132 (2025)
47. Udupa, S., Gurunath, P., Sikdar, A., Sundaram, S.: Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5904–5914 (2024)
48. Wei, Z., Chen, L., Jin, Y., Ma, X., Liu, T., Ling, P., Wang, B., Chen, H., Zheng, J.: Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28619–28630 (2024)
49. Wen, C., Peng, Z., Huang, Y., Yang, X., Shen, W.: Domain generalization in clip via learning with diverse text prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9559–9569 (2025)
50. Xu, F., Deng, S., Jia, T., Yu, X., Chen, D.: Ensembling disentangled domain-specific prompts for domain generalization. Knowledge-Based Systems **301**, 112358 (2024)
51. Xu, Q., Yao, L., Jiang, Z., Jiang, G., Chu, W., Han, W., Zhang, W., Wang, C., Tai, Y.: Dirl: Domain-invariant representation learning for generalizable semantic segmentation. In: Proceedings of the AAAI conference on artificial intelligence. pp. 2884–2892 (2022)
52. Xu, Z., Cheng, D., Jiang, X., Wang, N., Li, D., Gao, X.: Adversarial domain prompt tuning and generation for single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18584–18595 (2025)
53. Yang, S., Tian, Z., Jiang, L., Jia, J.: Unified language-driven zero-shot domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23407–23415 (2024)
54. Yang, Y., Deng, J., Li, W., Duan, L.: Resclip: Residual attention for training-free dense vision-language inference. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29968–29978 (2025)
55. Yi, J., Bi, Q., Zheng, H., Zhan, H., Ji, W., Huang, Y., Li, Y., Zheng, Y.: Learning spectral-decomposited tokens for domain generalized semantic segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8159–8168 (2024)

56. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
57. Yu, X., Yoo, S., Lin, Y.: Clipceil: Domain generalization through clip via channel refinement and image-text alignment. In: Advances in Neural Information Processing Systems. pp. 4267–4294 (2024)
58. Yun, S., Chae, S., Lee, D., Ro, Y.: Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25602–25612 (2025)
59. Zhang, X., Gu, S.S., Matsuo, Y., Iwasawa, Y.: Domain prompt learning for efficiently adapting clip to unseen domains. Transactions of the Japanese Society for Artificial Intelligence **38**(6), B–MC2_1 (2023)
60. Zhang, X., Tan, R.T.: Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14527–14537 (2025)
61. Zhang, Z., Wu, A., Han, Y.: Style evolving along chain-of-thought for unknown-domain object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14225–14234 (2025)
62. Zhao, D., Li, J., Wang, S., Wu, M., Zang, Q., Sebe, N., Zhong, Z.: Fishertune: Fisher-guided robust tuning of vision foundation models for domain generalized segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15043–15054 (2025)
63. Zhao, Y., Zhong, Z., Zhao, N., Sebe, N., Lee, G.H.: Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In: European conference on computer vision. pp. 535–552. Springer (2022)
64. Zhong, Z., Zhao, Y., Lee, G.H., Sebe, N.: Adversarial style augmentation for domain generalized urban-scene segmentation. In: Advances in neural information processing systems. pp. 338–350 (2022)
65. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
66. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)