VULNDG: DIAGNOSING AND CALIBRATING SHIFT-INDUCED SEMANTIC VULNERABILITIES FOR DOMAIN GENERALIZATION SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain-generalized semantic segmentation (DGSS) aims to learn from labeled source-domain data and generalize to unseen domains without accessing target data during training. A key challenge is that different semantic relations exhibit different vulnerability to domain shifts, yet such vulnerable relations cannot be identified from target-domain feedback under the strict source-only setting. To address this issue, we propose a vulnerability-aware DGSS framework that follows an exposure–diagnosis–calibration–retention paradigm. We first construct a language-guided feature-level spectral proxy shift, where text-conditioned modulation perturbs domain-sensitive feature amplitudes while preserving phase information, thereby exposing latent distribution-sensitive responses using source data only. By comparing clean and shifted predictions, we identify clean-correct-to-shifted-wrong transitions and quantify directed class-pair vulnerability to discover confused relations and their associated hard regions. We then introduce text-derived semantic anchors to selectively pull vulnerable features toward their correct semantics while separating them from diagnosed confusing categories. Finally, an exponential moving average (EMA) teacher provides cross-style consistency and vulnerability-focused retention to stabilize robustness learning. All diagnostic and optimization signals are constructed exclusively from source-domain data. Extensive experiments demonstrate that the proposed framework consistently improves cross-domain segmentation performance on unseen domains.

1 INTRODUCTION

Semantic segmentation is fundamental to autonomous driving, robotic perception, and scene understanding, yet models trained on a source distribution often degrade in unseen environments. Domain-generalized semantic segmentation (DGSS) therefore aims to generalize to unseen domains without accessing target-domain data during training. Existing approaches improve robustness through domain randomization, normalization and whitening, style diversification, category-aware memory, and consistency learning (Yue et al., 2019; Choi et al., 2021; Peng et al., 2022; Lee et al., 2022; Zhao et al., 2022; Kim et al., 2022; Zhong et al., 2022). More recently, vision foundation models (VFMs) and vision-language models (VLMs), such as DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021), have enabled stronger transferable representations for DGSS (Wei et al., 2024; Fahes et al., 2024; Benigmim et al., 2024; Pak et al., 2024; Zhang & Tan, 2025). Despite these advances, existing methods mainly focus on improving representations or diversifying training distributions, while the unequal vulnerability of semantic relations under domain shifts remains largely underexplored.

Frequency-domain modeling provides a natural way to characterize domain-sensitive variations. FDA (Yang & Soatto, 2020) manipulates low-frequency Fourier statistics at the image level, while SET (Yi et al., 2024) extends spectral decomposition to frozen VFM features. Rather than using spectral perturbation merely for augmentation or style invariance, we exploit it as a controllable source-only stress test: class relations that are stable on clean source data may become confused once domain-sensitive feature statistics are perturbed. This enables us to ask a more specific question: *which semantic relations are most vulnerable to domain shifts?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

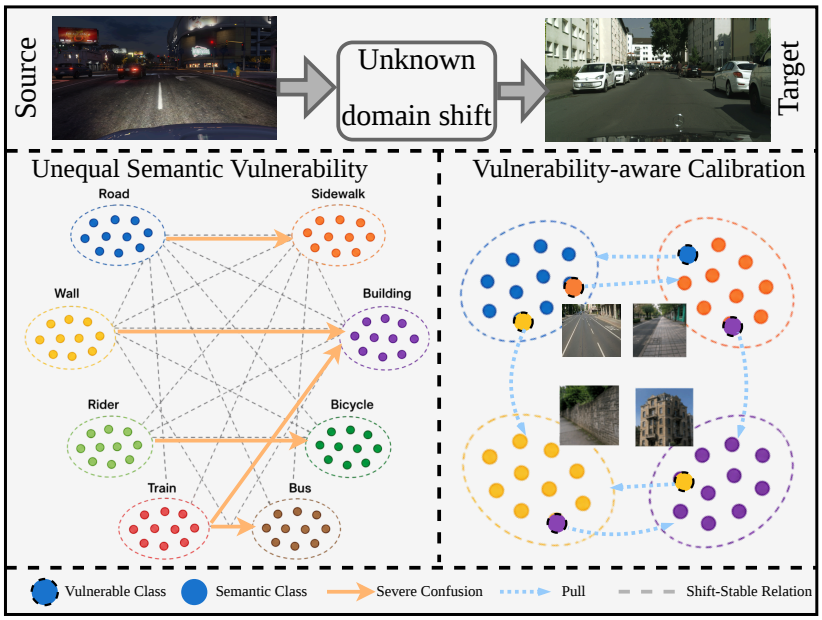

Figure 1: Motivation of our vulnerability-aware DGSS framework. Unseen domain shifts affect semantic relations unequally: some relations remain stable, whereas vulnerable classes exhibit directional confusion toward categories. This motivates us to diagnose vulnerable semantic relations and perform targeted calibration rather than imposing uniform robustness across all classes.

To address this question, we propose a vulnerability-aware DGSS framework following an *exposure–diagnosis–calibration–retention* paradigm. A language-guided feature-level spectral proxy shift first exposes latent shift-sensitive responses using source data only. We then diagnose directed vulnerable class relations from clean-to-shift prediction changes and identify their associated hard regions. The diagnosed vulnerabilities guide text-derived semantic-anchor calibration, while an exponential moving average (EMA) teacher (Tarvainen & Valpola, 2017; Hinton et al., 2015) provides cross-style consistency and robustness retention. Importantly, all diagnostic and optimization signals are constructed exclusively from source-domain data.

The main contributions are summarized as follows:

- We introduce a language-conditioned bounded spectral stress generator that exposes latent shift-sensitive semantic decisions using source data only.
- We propose a stable online vulnerability diagnosis mechanism that aggregates clean-correct-to-shifted-wrong transitions to discover directed vulnerable class relations and their hard regions without manually specified confusing categories.
- We develop vulnerability-aware semantic calibration and EMA-guided robustness retention, and empirically verify that the source-diagnosed vulnerabilities are predictive of semantic confusion under real unseen-domain shifts.

2 Related Work

2.1 Domain Generalized Semantic Segmentation

Domain generalized semantic segmentation (DGSS) aims to learn from labeled source data and generalize to unseen domains without target-domain access. Early studies improve generalization through domain randomization, feature normalization, style diversification, and domain-invariant representation learning (Yue et al., 2019; Zhou et al., 2021; Tang et al., 2021; Choi et al., 2021). Subsequent methods further exploit semantic-aware normalization, content/style diversification, consistency learning, adversarial style augmentation, and class-level memory (Peng et al., 2022; Lee et al.,

2022; Zhao et al., 2022; Zhong et al., 2022; Kim et al., 2022). Transformer-based approaches such as HGFormer (Ding et al., 2023) and IBAFormer (Sun et al., 2023) further explore hierarchical grouping and inter-sample interactions. Recent surveys summarize these developments and highlight the growing importance of foundation models in DGSS (Rafi et al., 2024; Schwonberg & Gottschalk, 2025). Different from these approaches, we explicitly diagnose the unequal shift-induced vulnerability of semantic relations and perform targeted source-only calibration.

2.2 Foundation Models and Language Guidance for DGSS

Large-scale pretrained VFMs and VLMs, such as DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021), have recently provided stronger transferable priors for DGSS. Rein (Wei et al., 2024) efficiently adapts frozen VFMs, while VLTSeg (Hümmer et al., 2023) demonstrates the strong cross-domain generalization of CLIP-based transfer learning. Language knowledge is further exploited for style diversification in FAMix (Fahes et al., 2024) as well as in language-guided semantic alignment and unseen-domain extension (Jiang et al., 2025).tqdm (Pak et al., 2024) further employs language-derived domain-invariant object queries. CLOUDS (Benigmim et al., 2024) combines CLIP, generative models, and SAM for complementary generalization cues, QueryDiff (Li et al., 2025) leverages diffusion-model guidance to learn generalizable semantic representations, whereas MFuser (Zhang & Tan, 2025) explicitly integrates VFM and VLM representations. In contrast, we exploit language for both spectral shift construction and semantic calibration, enabling source-only diagnosis of vulnerable class relations.

2.3 Spectral Modeling and Robustness Regularization

Frequency-domain modeling has been widely used to characterize domain-sensitive appearance statistics (Huang et al., 2021; Xu et al., 2021). FDA (Yang & Soatto, 2020) manipulates low-frequency Fourier components at the image level, while SET (Yi et al., 2024) extends spectral decomposition to frozen VFM features by separately modeling amplitude and phase. Different from these methods, we employ language-conditioned feature-amplitude perturbation as a source-only stress test to expose semantic vulnerability rather than merely improve style invariance. Meanwhile, teacher–student learning and EMA models provide effective mechanisms for stable knowledge transfer (Tarvainen & Valpola, 2017; Hinton et al., 2015). We leverage these principles for cross-style consistency and selectively retain historical knowledge in source-diagnosed vulnerable regions.

3 Method

3.1 Problem Definition and Method Overview

Given a labeled source-domain dataset $\mathcal{D}_s = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{N}$, domain-generalized semantic segmentation (DGSS) aims to learn a segmentation model that generalizes to unseen target domains. We consider a strict source-only setting, where no target-domain images, labels, predictions, evaluation results, or statistics are accessible during training or hyperparameter selection.

Our key observation is that semantic relations exhibit unequal vulnerability to domain shifts. A class relation that is reliably separated on clean source data may become confused once domain-sensitive feature statistics change. Accordingly, VulnDG follows an *Expose–Diagnose–Calibrate–Retain* paradigm. First, a language-guided spectral proxy shift exposes latent shift-sensitive responses using source data alone. We then identify directed semantic failures from clean-correct to shifted-wrong predictions and use them to diagnose vulnerable class relations and pixels. These diagnosed vulnerabilities guide targeted semantic-anchor calibration, while an EMA teacher preserves stable semantic knowledge during robustness optimization.

During training, the image encoder E and text encoder T remain frozen. The segmentation student consists of a shared learnable projector g_{ϕ_S} and segmentation predictor H_{θ_S}, with parameters $\Theta_S = \{\phi_S, \theta_S\}$. A lightweight text-to-spectrum mapper M_ψ generates language-conditioned spectral modulation parameters. The clean and proxy-shifted branches share the same student prediction path. An EMA teacher maintains a temporally smoothed copy of Θ_S and receives no gradient. The proxy-shift generator and EMA teacher are used only during training and discarded at inference.

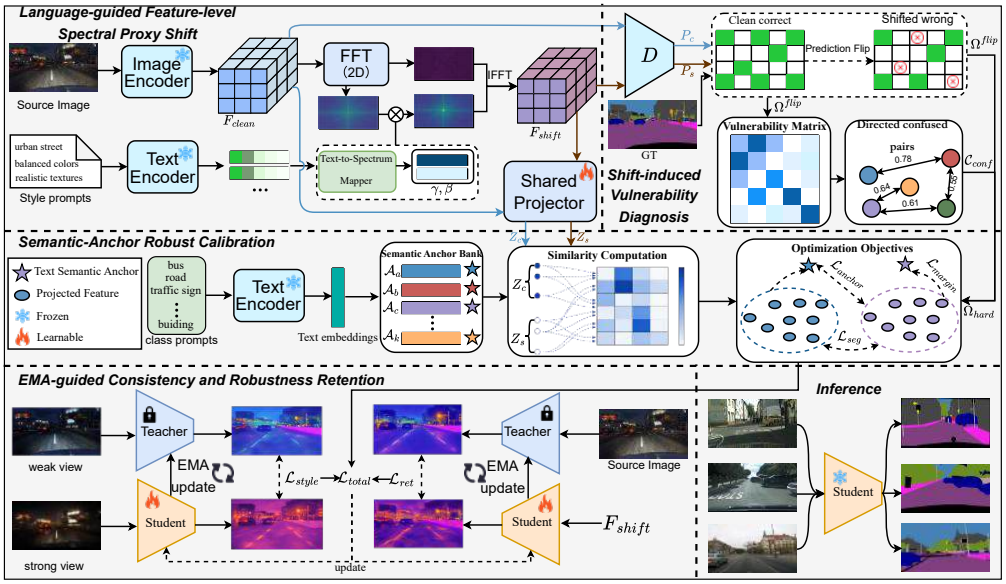

Figure 2: Overview of VulnDG. A language-guided feature-level spectral proxy shift first exposes shift-sensitive semantic responses from source data. Clean-correct-to-shifted-wrong transitions are then exploited to diagnose directed vulnerable class relations and their associated hard regions. The diagnosed vulnerabilities guide semantic-anchor calibration, while an EMA teacher provides cross-style consistency and historical retention. All training signals are derived exclusively from the source domain, and only the trained student is used for inference.

3.2 Language-guided Feature-level Spectral Proxy Shift

The spectral proxy shift acts as a source-internal stress test. Rather than attempting to reproduce a particular target-domain appearance, it perturbs domain-sensitive feature statistics to expose semantic decisions that are potentially unstable under appearance changes. For a source image $\mathbf{x}_s$, the frozen image encoder extracts $\mathbf{F}_s^{\text{clean}} = E(\mathbf{x}_s)$. The clean representation and prediction are $\mathbf{Z}_s^{\text{clean}} = g_{\phi_S}(\mathbf{F}_s^{\text{clean}})$ and $\mathbf{P}_s^{\text{clean}} = \text{softmax}(H_{\theta_S}(\mathbf{Z}_s^{\text{clean}}))$, respectively. The clean branch is supervised by $\mathcal{L}_{\text{seg}} = \ell_{\text{seg}}(H_{\theta_S}(\mathbf{Z}_s^{\text{clean}}), \mathbf{y}_s)$, where ℓ_{seg} denotes the native segmentation objective of the underlying architecture. We first apply a 2D Fourier transform along the spatial dimensions:

$$\mathcal{F}(\mathbf{F}_s^{\text{clean}}) = \mathbf{A}_s^F \odot \exp\left(j\mathbf{\Phi}_s^F\right), \tag{1}$$

where $\mathbf{A}_s^F$ and $\mathbf{\Phi}_s^F$ denote the amplitude and phase spectra, respectively. To condition the perturbation on language, we define a fixed set of generic style prompts $\mathcal{P}_{\text{sty}}$. For a sampled prompt $p_k^{\text{sty}} \in \mathcal{P}_{\text{sty}}$, the frozen text encoder produces $\mathbf{u}_k^{\text{sty}} = T(p_k^{\text{sty}})/\|T(p_k^{\text{sty}})\|_2$. The text-to-spectrum mapper converts this embedding into channel-wise modulation parameters:

$$(\tilde{\boldsymbol{\gamma}}_k, \tilde{\boldsymbol{\beta}}_k) = M_\psi(\mathbf{u}_k^{\text{sty}}), \quad \boldsymbol{\gamma}_k = \alpha \tanh(\tilde{\boldsymbol{\gamma}}_k), \quad \boldsymbol{\beta}_k = \alpha \tanh(\tilde{\boldsymbol{\beta}}_k), \tag{2}$$

where α bounds the overall modulation strength. The modulation is injected only into a centered low-frequency region:

$$\mathbf{A}_s^{F,\text{shift}} = \mathbf{A}_s^F + \mathbf{M}_F \odot \left[\text{Softplus}\left((1+\boldsymbol{\gamma}_k) \odot \mathbf{A}_s^F + \boldsymbol{\beta}_k\right) - \mathbf{A}_s^F\right], \tag{3}$$

where $\mathbf{M}_F$ denotes the low-frequency mask. Frequencies outside the selected region remain unchanged, while Softplus keeps the modulated amplitude non-negative. The original phase is retained when reconstructing the shifted feature:

$$\begin{aligned} \mathbf{F}_s^{\text{shift}} &= \mathcal{F}^{-1}\left(\mathbf{A}_s^{F,\text{shift}} \odot \exp(j\mathbf{\Phi}_s^F)\right), \\ \mathbf{P}_s^{\text{shift}} &= \text{softmax}\left(H_{\theta_S}\left(g_{\phi_S}(\mathbf{F}_s^{\text{shift}})\right)\right). \end{aligned} \tag{4}$$

Preserving the original phase encourages the proxy shift to modify appearance-sensitive statistics without changing the underlying spatial semantic layout. Importantly, both clean and shifted features are processed by the same projector and segmentation predictor. The shifted branch is supervised using the same source annotation through $\mathcal{L}_{\text{shift}} = \ell_{\text{seg}}(H_{\theta_S}(g_{\phi_S}(\mathbf{F}_s^{\text{shift}})), \mathbf{y}_s)$.

3.3 Shift-induced Vulnerability Diagnosis

The proxy shift provides a controlled way to test whether a semantic decision that is correct on the clean source representation remains stable after domain-sensitive statistics are perturbed. We therefore focus on strict *clean-correct/shifted-wrong* transitions rather than ordinary source-domain confusion. Let $\hat{y}_i^{\text{clean}} = \arg\max_c P_{s,i,c}^{\text{clean}}$ and $\hat{y}_i^{\text{shift}} = \arg\max_c P_{s,i,c}^{\text{shift}}$. For a directed semantic transition from class a to class b, we define

$$\Omega_{a\to b}^{\text{flip}} = \left\{ i \mid y_i = a,\ \hat{y}_i^{\text{clean}} = a,\ \hat{y}_i^{\text{shift}} = b,\ a \neq b \right\}. \tag{5}$$

The source ground truth is used only to establish a reliable semantic reference: it identifies predictions known to be correct before perturbation and determines the source class a. Thus, the diagnosis isolates failures induced by the proxy shift instead of mixing them with ordinary source-domain errors. Let $\Omega_a^{\text{correct}} = \{i \mid y_i = a, \hat{y}_i^{\text{clean}} = a\}$. The directed flip frequency is

$$\rho_{a,b} = \frac{|\Omega_{a\to b}^{\text{flip}}|}{|\Omega_a^{\text{correct}}| + \epsilon}. \tag{6}$$

This quantity measures how frequently a clean-correct class-a prediction drifts toward class b under the proxy shift. Frequency alone does not describe the magnitude of the induced semantic change. We therefore measure transition severity as

$$\delta_{a,b} = \frac{\sum_{i \in \Omega_{a\to b}^{\text{flip}}} D_{\text{JS}}\left(\mathbf{P}_{s,i}^{\text{clean}} \| \mathbf{P}_{s,i}^{\text{shift}}\right)}{|\Omega_{a\to b}^{\text{flip}}| + \epsilon}, \tag{7}$$

where D_{JS} denotes Jensen–Shannon divergence. The final directed vulnerability score is

$$V_{a,b} = \rho_{a,b}\delta_{a,b}. \tag{8}$$

The two factors capture complementary properties: $\rho_{a,b}$ estimates how often the failure occurs, while $\delta_{a,b}$ quantifies its conditional severity. Up to the numerical stabilization terms, their product corresponds to the accumulated semantic deviation of the $a \to b$ failures normalized by the clean-correct support of class a. Hence, a high vulnerability score requires a transition to be both recurrent and substantial. Collecting all scores gives a directed vulnerability matrix $\mathbf{V} \in \mathbb{R}^{C\times C}$. To suppress unreliable statistics, only classes satisfying $|\Omega_a^{\text{correct}}| \geq N_{\min}$ participate in pair mining. For each eligible class, we retain the K largest off-diagonal entries:

$$\mathcal{C}_{\text{conf}} = \bigcup_{\substack{a \\ |\Omega_a^{\text{correct}}| \geq N_{\min}}} \left\{ (a,b) \mid b \in \text{TopK}_{b\neq a}(\mathbf{V}_{a,:}) \right\}. \tag{9}$$

Since mining is performed row-wise, $a \to b$ and $b \to a$ remain distinct relations. The spatial regions associated with the selected relations are further collected as

$$\Omega_{\text{hard}} = \bigcup_{(a,b)\in\mathcal{C}_{\text{conf}}} \Omega_{a\to b}^{\text{flip}}. \tag{10}$$

Thus, $\mathcal{C}_{\text{conf}}$ specifies *which semantic relations* are vulnerable, whereas Ω_{hard} determines *where* targeted calibration and retention should be applied.

3.4 Semantic-Anchor Robust Calibration

The diagnosis above identifies semantic relations that become unstable under the proxy shift. We next exploit frozen text semantics to explicitly correct these relations without modifying the text anchors themselves.

For class c, the frozen text encoder produces the normalized semantic anchor $\mathbf{e}_c = T(t_c)/\|T(t_c)\|_2$. For a projected pixel representation, we similarly use $\mathbf{z}_i^v = \mathbf{Z}_{s,i}^v/\|\mathbf{Z}_{s,i}^v\|_2$, where $v \in \{\text{clean}, \text{shift}\}$. Their semantic relation is represented by

$$s_{i,c}^v = \cos(\mathbf{z}_i^v, \mathbf{e}_c), \quad \mathbf{q}_i^v = \text{softmax}\left([s_{i,1}^v, \ldots, s_{i,C}^v]\right). \tag{11}$$

We adopt cosine similarity because both visual representations and text anchors are L_2-normalized and therefore lie on a shared unit hypersphere. Under this normalization, squared Euclidean distance induces the same similarity ordering, whereas cosine similarity introduces no additional metric parameters. The anchor-induced distribution $\mathbf{q}_i^v$ is used only for semantic calibration and is distinct from the segmentation probability $\mathbf{P}$. We first preserve the clean semantic relation under the proxy shift:

$$\mathcal{L}_{\text{sem-cons}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} D_{\text{KL}}\left(\text{sg}[\mathbf{q}_i^{\text{clean}}] \| \mathbf{q}_i^{\text{shift}}\right), \tag{12}$$

where $\text{sg}[\cdot]$ stops gradients through the clean reference. The one-way constraint prevents the clean and shifted semantic distributions from drifting together. Semantic consistency alone does not guarantee that a vulnerable shifted feature moves sufficiently toward its correct semantic anchor. We therefore introduce

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\Omega_{\text{hard}}| + \epsilon} \sum_{i \in \Omega_{\text{hard}}} \left(1 - s_{i,y_i}^{\text{shift}}\right), \tag{13}$$

which explicitly attracts diagnosed vulnerable features toward their ground-truth semantic anchors.

For each diagnosed relation $(a, b) \in \mathcal{C}_{\text{conf}}$, we further encourage the correct anchor to remain separated from its confusing category:

$$\mathcal{L}_{\text{margin}} = \frac{\sum_{(a,b) \in \mathcal{C}_{\text{conf}}} \sum_{i \in \Omega_{a \to b}^{\text{flip}}} \left[m - \left(s_{i,a}^{\text{shift}} - s_{i,b}^{\text{shift}}\right)\right]_+}{|\Omega_{\text{hard}}| + \epsilon}, \tag{14}$$

where $m > 0$ denotes the desired semantic margin and $[z]_+ = \max(0, z)$. The complete vulnerability-aware calibration objective is

$$\mathcal{L}_{\text{cal}} = \mathcal{L}_{\text{sem-cons}} + \mathcal{L}_{\text{anchor}} + \mathcal{L}_{\text{margin}}. \tag{15}$$

The three terms respectively preserve clean semantic relations, pull vulnerable representations toward their correct anchors, and explicitly separate them from the diagnosed confusing categories.

3.5 EMA-guided Consistency and Robustness Retention

Although vulnerability-aware calibration strengthens unstable semantic relations, robustness-oriented optimization may also disturb predictions that are already reliable. We therefore maintain an EMA teacher as a temporally smoothed reference. Since the image encoder remains frozen, the teacher tracks only the trainable prediction-path parameters. It is initialized from the student and updated after each optimization step as

$$\Theta_T^{(t+1)} = \mu \Theta_T^{(t)} + (1 - \mu)\Theta_S^{(t+1)}, \tag{16}$$

where μ is the EMA momentum. The teacher receives no gradient.

Cross-style consistency. Weak and strong appearance-perturbed views are independently sampled from the same source image as $\mathbf{x}^w = \mathcal{A}(\mathbf{x}_s; \xi_w)$ and $\mathbf{x}^{st} = \mathcal{A}(\mathbf{x}_s; \xi_{st})$, where the strong branch spans a wider appearance range while preserving geometry. Their predictions are $\mathbf{R}_T^w = \text{sg}[\text{softmax}(f_{\Theta_T}(\mathbf{x}^w))]$ and $\mathbf{R}_S^{st} = \text{softmax}(f_{\Theta_S}(\mathbf{x}^{st}))$. Reliable teacher pixels are selected by $M_i = \mathbf{1}[\max_c R_{T,i,c}^w > \tau_c]$. The cross-style consistency objective is

$$\mathcal{L}_{\text{style}} = \frac{\sum_{i \in \Omega} M_i D_{\text{KL}}\left(\mathbf{R}_{T,i}^w \| \mathbf{R}_{S,i}^{st}\right)}{\sum_{i \in \Omega} M_i + \epsilon}. \tag{17}$$

The weak EMA-teacher view provides a relatively stable semantic target, whereas the strongly perturbed student is explicitly optimized to preserve the same prediction under larger source-side appearance variation.

Vulnerability-focused historical retention. Cross-style consistency acts broadly over reliable source pixels, while regions exposed as vulnerable require more targeted preservation. The clean EMA-teacher distribution is $\mathbf{R}_T^{\text{clean}} = \text{sg}[\text{softmax}(f_{\Theta_T}(\mathbf{x}_s))]$. Since $\mathbf{F}_s^{\text{shift}}$ is already encoded, it directly enters the student prediction path to obtain $\mathbf{R}_S^{\text{shift}} = \text{softmax}(H_{\theta_S}(g_{\phi_S}(\mathbf{F}_s^{\text{shift}})))$. Historical retention is applied only to the diagnosed hard region:

$$\mathcal{L}_{\text{ret}} = \frac{1}{|\Omega_{\text{hard}}| + \epsilon} \sum_{i \in \Omega_{\text{hard}}} D_{\text{KL}} \left(\mathbf{R}_{T,i}^{\text{clean}} \| \mathbf{R}_{S,i}^{\text{shift}} \right). \tag{18}$$

The clean-to-shifted asymmetry is intentional. The clean EMA prediction provides a temporally stable semantic reference, whereas the shifted student representation is precisely the representation exposed as vulnerable. Applying the same proxy shift to the reference branch would weaken its role as a stable historical target.

3.6 Overall Optimization

The clean and proxy-shifted branches are supervised using the same source-domain annotations, giving $\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{shift}}$. The complete student objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cal}} \mathcal{L}_{\text{cal}} + \lambda_{\text{ema}} \left(w(t) \mathcal{L}_{\text{style}} + \mathcal{L}_{\text{ret}} \right), \tag{19}$$

where $w(t) = \min(1, 10t/T)$ linearly activates cross-style consistency during the first 10% of training. This warm-up avoids imposing an overly strong consistency constraint from an early EMA teacher whose predictions are not yet sufficiently stable. λ_{cal} controls vulnerability-aware semantic calibration, whereas λ_{ema} controls EMA-guided consistency and robustness retention.

All method-specific settings are determined exclusively from source-domain data and remain fixed across all unseen domains. No target-domain image, prediction, statistic, evaluation result, or validation performance is used for vulnerability mining, hyperparameter selection, checkpoint selection, or stopping criteria.

4 Experiments

Network Architecture We adopt DINOv2-L (Oquab et al., 2023) and SIGLIP-L (Zhai et al., 2023) as the VFM and VLM, respectively, with Mask2Former (Cheng et al., 2022) as the segmentation decoder following MFuser (Zhang & Tan, 2025). The pretrained encoders are frozen, while the projector and decoder are optimized. An EMA teacher is used only during training, and the trained student is retained for inference.

Implementation Details. We conduct experiments on GTA5 (Richter et al., 2016), Cityscapes (Cordts et al., 2016), BDD100K (Yu et al., 2020), Mapillary (Neuhold et al., 2017), and ACDC (Sakaridis et al., 2021), covering synthetic-to-real, real-to-real, and clear-to-adverse-weather generalization settings. We use an input size of 512×512 and a batch size of 2. The model is optimized using AdamW with an initial learning rate of 1×10^{-4} and trained for 20k iterations. The pretrained DINOv2-L and SIGLIP-L encoders are kept frozen, while the newly introduced components and segmentation decoder are optimized. All method-specific hyperparameters are fixed across target domains and determined solely from source-domain data.

4.1 Comparison with State-of-The-Art Methods

We compare our method with existing DGSS methods on two setups:synthetic-to-real(G$\rightarrow$ {C,B,M}) and real-to-real(C$\rightarrow$ {B,M}). We mainly compare with recent foundation model-based approaches, including CLOUDS (Benigmim et al., 2024), VLTseg (Hümmer et al., 2023),Rein (Wei et al., 2024), SET (Yi et al., 2024), tqdm (Pak et al., 2024) and Mfuser (Zhang & Tan, 2025).

As shown in Table 1, VulnDG consistently achieves the best performance under both synthetic-to-real and real-to-real domain generalization settings. For $G \rightarrow \{C, B, M\}$, VulnDG reaches 72.81%, 63.50%, and 72.97% mIoU, respectively, with the highest average of 69.76%. Compared with MFuser under the same DINOv2-L and SIGLIP-L configuration, the gains are 1.71, 2.31, and 1.26 points on the three target domains, yielding an average improvement of 1.76 points. The largest

Table 1: Performance comparison (mIoU, %) under synthetic-to-real ($G \rightarrow \{C, B, M\}$) and real-to-real ($C \rightarrow \{B, M\}$) domain generalization settings. DINOv2 L (Oquab et al., 2023) is used as the VFM for MFuser and VulnDG, with SIGLIP-L as the VLM. Best results are shown in bold.

Method	Backbone	Synthetic-to-Real				Real-to-Real		
		$G \rightarrow C$	$G \rightarrow B$	$G \rightarrow M$	Avg.	$C \rightarrow B$	$C \rightarrow M$	Avg.
SAN-SAW	RN101	45.33	41.18	40.77	42.43	54.73	61.27	58.00
WildNet	RN101	45.79	41.73	47.08	44.87	47.01	50.94	48.98
SHADE	RN101	46.66	43.66	45.50	45.27	50.95	60.67	55.81
TLDR	RN101	47.58	44.88	48.80	47.09	–	–	–
FAMix	RN101	49.47	46.40	51.97	49.28	–	–	–
SHADE	MiT-B5	53.27	48.19	54.99	52.15	–	–	–
IBAFormer	MiT-B5	56.34	49.76	58.26	54.79	–	–	–
VLTSeg	CLIP-B	47.50	45.70	54.30	49.17	–	–	–
HGFormer	Swin-L	–	–	–	–	61.50	72.10	66.80
CLOUDS	ConvNeXt-L	60.20	57.40	67.00	61.50	–	–	–
VLTSeg	EVA02-L	65.60	58.40	66.50	63.50	64.40	76.40	70.40
Rein	EVA02-L	65.30	60.50	64.90	63.60	64.10	69.50	66.80
Rein	DINOv2-L	66.40	60.40	66.10	64.30	65.00	72.30	68.65
SET	DINOv2-L	68.06	61.64	67.68	65.79	65.07	75.67	70.37
tqdm	EVA02-L	68.88	59.18	70.10	66.05	64.72	76.15	70.44
MFuser	SIGLIP-L	71.10	61.19	71.71	68.00	65.58	78.10	71.84
VulnDG	SIGLIP-L	**72.81**	**63.50**	**72.97**	**69.76**	**66.42**	**78.91**	**72.67**

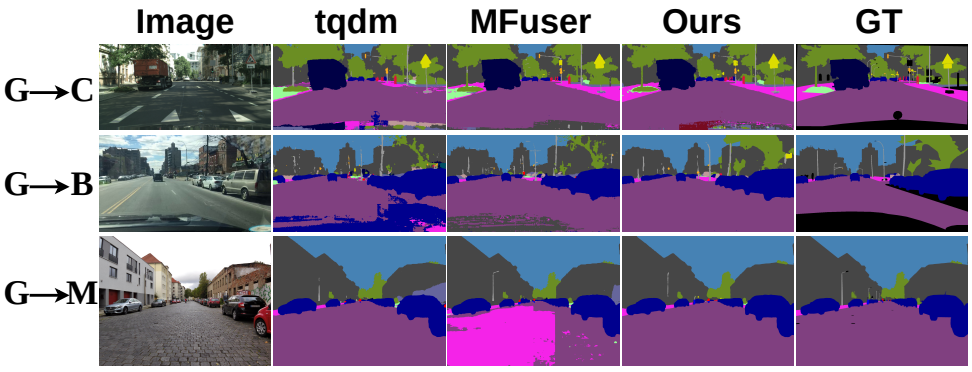

Figure 3: Qualitative visualizations on target domains under the G $\rightarrow$ {C,B,M} settings. VulnDG is compared with tqdm (Pak et al., 2024) and MFuser (Zhang & Tan, 2025)

gain is observed on BDD100K, suggesting that explicitly diagnosing and calibrating shift-sensitive semantic relations is particularly beneficial when the source-to-target discrepancy is more challenging. The improvement remains consistent in the real-to-real setting: VulnDG improves BDD100K and Mapillary by 0.84 and 0.81 points over MFuser, increasing the average mIoU from 71.84% to 72.67%. Overall, the consistent gains across synthetic and real source domains demonstrate that VulnDG improves cross-domain robustness without relying on a specific source-domain distribution.

4.1.1 Clear-to-Adverse-Weather Generalization

As shown in Table 2, VulnDG achieves the best performance across all four adverse weather conditions, obtaining 83.37%, 59.23%, 79.15%, and 75.74% mIoU under Fog, Night, Rain, and Snow, respectively, with the highest average mIoU of 74.37%. Compared with MFuser, VulnDG improves the performance by 2.33, 3.26, 5.89, and 0.84 points across the four conditions, yielding an average gain of 3.08 points. Notably, the largest improvement is observed under the challenging Night

Table 2: Performance comparison (mIoU in %) under the clear-to-adverse-weather setting (C→{Fog, Night, Rain, Snow}). Our method is marked in gray. The best results for each condition and the average performance are highlighted in bold.

Method	Backbone	clear-to-adverse-weather				
		Fog	*Night*	*Rain*	*Snow*	Avg.
IBN	RN50	63.80	21.20	50.40	49.60	46.25
IW	RN50	62.40	21.80	52.40	47.60	46.05
ISW	RN50	64.30	24.30	54.00	49.80	48.60
ISSA	MiT-B5	67.50	33.20	55.90	53.20	52.45
CMFormer	Swin-L	77.80	33.70	67.60	64.30	60.85
Rein	DINOv2-L	79.48	55.92	72.45	70.57	69.61
SET	DINOv2-L	80.06	57.29	74.80	73.69	71.46
tqdm	EVA02-L	81.28	54.80	72.92	72.41	70.35
MFuser	SIGLIP	81.04	55.97	73.26	74.90	71.29
VulnDG	SIGLIP	**83.37**	**59.23**	**79.15**	**75.74**	**74.37**

Table 3: Ablation on different VFM and VLM combinations under the synthetic-to-real (G→{C,B,M}) setting.

VFM + VLM	C	B	M	Avg.
DINOv2 + CLIP	71.92	62.88	72.21	69.00
DINOv2 + SIGLIP	72.81	63.50	72.97	69.76
DINOv2 + EVA02	72.07	63.12	72.56	69.25
DINOv3 + CLIP	72.41	62.74	72.96	69.37
DINOv3 + SIGLIP	72.63	64.12	73.41	70.05
DINOv3 + EVA02	71.96	63.68	73.18	69.61

Table 4: Component ablation study of VulnDG under the synthetic-to-real setting.

Configuration	C	B	M	Avg.
Baseline	71.10	61.19	71.71	68.00
+ Spectral Proxy Shift	71.62	61.73	72.08	68.48
+ Semantic Consistency	71.91	62.05	72.36	68.77
+ Vulnerability Calibration	72.28	62.62	72.64	69.18
+ EMA Cross-style Consistency	72.51	63.02	72.86	69.46
+ Robustness Retention	72.81	63.50	72.97	69.76

condition, indicating stronger robustness to severe illumination and appearance shifts. The consistent gains across diverse adverse conditions further demonstrate the effectiveness of diagnosing and calibrating shift-induced semantic vulnerabilities for robust generalization under challenging environmental changes.

4.1.2 Impact of VFM–VLM Pairing

As shown in Table 3, DINOv3 generally improves the average performance over DINOv2 across different VLMs, although the gains vary across target domains. SIGLIP consistently performs best with both VFMs, and DINOv3+SIGLIP achieves the highest average mIoU of 70.05%. Notably, DINOv2+SIGLIP remains slightly better on Cityscapes, suggesting that stronger visual representations do not uniformly benefit every domain and that VFM–VLM complementarity is important for cross-domain generalization.

4.1.3 Component Ablation

As shown in Table 4, progressively introducing each component consistently improves cross-domain performance. The spectral proxy shift and semantic consistency enhance robustness to distribution variations, while vulnerability-guided calibration further improves the average mIoU to 69.18%. Incorporating EMA-based consistency and retention finally achieves 69.76%, yielding a 1.76-point improvement over the baseline. These results demonstrate the complementary effectiveness of the proposed components.

5 Conclusion

We propose a vulnerability-aware source-only DGSS framework following an exposure–diagnosis–calibration–retention paradigm. By using controllable spectral proxy shifts, the proposed method explicitly identifies shift-sensitive semantic relations and performs targeted semantic calibration, while EMA-based consistency and retention further stabilize robust knowledge across source-side variations. Extensive experiments across different settings demonstrate the effectiveness and robustness of our approach for generalizing to diverse unseen domains.

AI USE STATEMENT

Generative AI tools were used solely to improve the grammar, wording, clarity, and readability of the manuscript.All AI-assisted edits were carefully reviewed and approved by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. 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.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 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.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 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.

Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. 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.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. 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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6891–6902, 2021.

Christoph Hümmer, Manuel Schwonberg, Liangwei Zhou, Hu Cao, Alois Knoll, and Hanno Gottschalk. Vltseg: Simple transfer of clip-based vision-language representations for domain generalized semantic segmentation. In *ACCV*, 2023.

Chenyi Jiang, Jianqin Zhao, Jingjing Deng, Zechao Li, and Haofeng Zhang. Imbuing, enrichment and calibration: Leveraging language for unseen domain extension. *International Journal of Computer Vision*, 133(7):4064–4090, 2025.

Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. 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.

Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. 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.

Fan Li, Xuan Wang, Min Qi, Zhaoxiang Zhang, and Yuelei Xu. Better to teach than to give: Domain generalized semantic segmentation via agent queries with diffusion model guidance. In *Forty-second International Conference on Machine Learning*, 2025.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. In *European Conference on Computer Vision*, pp. 37–54. Springer, 2024.

Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2594–2605, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Taki Hasan Rafi, Ratul Mahjabin, Emon Ghosh, Young-Woong Ko, and Jeong-Gun Lee. Domain generalization for semantic segmentation: a survey. *Artificial Intelligence Review*, 57(9):1, 2024.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision*, pp. 102–118, 2016.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10765–10775, 2021.

Manuel Schwonberg and Hanno Gottschalk. Domain generalization for semantic segmentation: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6437–6448, 2025.

Qiyu Sun, Huilin Chen, Meng Zheng, Ziyan Wu, Michael Felsberg, and Yang Tang. Ibaformer: Intra-batch attention transformer for domain generalized semantic segmentation. *arXiv preprint arXiv:2309.06282*, 2023.

Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, and Dimitris N Metaxas. Crossnorm and selfnorm for generalization under distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 52–61, 2021.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. 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.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14383–14392, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4085–4095, 2020.

Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning spectral-decomposited tokens for domain generalized semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8159–8168, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 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.

Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2100–2110, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Xin Zhang and Robby T Tan. 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.

Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European conference on computer vision*, pp. 535–552. Springer, 2022.

Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in neural information processing systems*, 35:338–350, 2022.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

A Additional Implementation Details

In addition to the datasets reported in the main paper, we further conduct experiments on SYNTHIA to evaluate the generalization ability of VulnDG under an additional synthetic-to-real setting. SYNTHIA provides diverse synthetic urban scenes with substantial variations in illumination, weather, viewpoint, and scene layout. We map its semantic annotations to the shared evaluation label space following the standard protocol. All training settings and method-specific hyperparameters are kept identical to those used in the main experiments, and no unseen-domain data or evaluation result is used for hyperparameter selection.

All experiments are implemented in PyTorch and conducted on NVIDIA TITAN GPUs. The default loss weights are set to $\lambda_{\text{cal}} = 1.00$ and $\lambda_{\text{ema}} = 0.50$ for all main experiments. Here, λ_{cal} controls the vulnerability-aware semantic calibration objective, whereas λ_{ema} balances the EMA-guided cross-style consistency and robustness retention terms. The same values are fixed across all source and unseen-domain combinations.

B Evaluate on SYNTHIA Benchmarks

As shown in Table 5 and Table 6,VulnDG consistently improves generalization performance when Synthia is used as either the source domain or an unseen target domain. Under the S→{C, B, M} setting, VulnDG achieves 54.05%, 46.74%, and 53.49% mIoU on Cityscapes, BDD100K, and Mapillary, respectively, yielding the best average performance of 51.43%. Compared with MFuser using the same SIGLIP-L VLM, the improvements are 1.85, 0.77, and 2.18 points on the three target domains, respectively, resulting in an average gain of 1.60 points.

When Synthia is instead treated as the unseen target domain, VulnDG also outperforms MFuser by 0.70 points under G→Synthia and by 0.52 points under C→Synthia. These consistent gains across different source–target configurations indicate that VulnDG is not tied to a particular source distribution or target-domain appearance. Overall, the results further demonstrate that diagnosing and calibrating shift-induced semantic vulnerabilities improves robustness across diverse synthetic and real-world domain shifts.

Table 5: Performance comparison (mIoU in %) under the synthetic-to-real setting (S→{C, B, M}).

Method	Backbone	**synthetic-to-real**			
		$S \rightarrow C$	$S \rightarrow B$	$S \rightarrow M$	Avg.
SAN-SAW	RN101	40.87	35.98	37.26	38.04
TLDR	RN101	42.60	35.46	37.46	38.51
IBAFormer	MiT-B5	50.92	44.66	50.58	48.72
Rein	DINOv2-L	48.59	44.42	48.64	47.22
SET	DINOv2-L	49.65	45.45	49.45	48.18
MFuser	SIGLIP-L	52.20	45.97	51.31	49.83
VulnDG	SIGLIP-L	**54.05**	**46.74**	**53.49**	**51.43**

Table 6: Performance comparison (mIoU in %) on Synthia under the G→Synthia and C→Synthia settings.

Method	Backbone	**G → Synthia**	**C → Synthia**
Rein	DINOv2-L	48.86	48.56
SET	DINOv2-L	50.01	49.61
MFuser	SIGLIP-L	52.02	51.64
VulnDG	SIGLIP-L	**52.72**	**52.16**

C Hyperparameter Sensitivity

We investigate the joint sensitivity of λ_{cal} and λ_{ema} by varying both weights over a broad range. As shown in Table 7, VulnDG maintains consistently strong performance across all nine configurations. The average mIoU varies only from 68.95% to 69.76%, with a standard deviation of approximately 0.24, indicating that the proposed method is not highly sensitive to the precise choice of the two loss weights. Similar stability is observed on the individual target domains, with performance ranges of 0.81, 1.05, and 0.62 points on Cityscapes, BDD100K, and Mapillary, respectively.

Table 7: Detailed sensitivity analysis of λ_{cal} and λ_{ema} under the synthetic-to-real setting.

λ_{cal}	λ_{ema}	$G \to C$	$G \to B$	$G \to M$	Avg.
0.50	0.25	72.05	62.45	72.42	68.97
0.50	0.50	72.20	62.72	72.55	69.16
0.50	1.00	72.00	62.50	72.35	68.95
1.00	0.25	72.42	63.02	72.70	69.38
1.00	0.50	72.81	63.50	72.97	69.76
1.00	1.00	72.48	63.06	72.72	69.42
2.00	0.25	72.14	62.70	72.46	69.10
2.00	0.50	72.44	63.00	72.76	69.40
2.00	1.00	72.19	62.78	72.55	69.17

D Computational Cost and Statistical Stability

As shown in Table 8, VulnDG achieves a favorable trade-off between model efficiency and cross-domain generalization. Compared with MFuser, VulnDG slightly reduces the total parameter count from 1127.5M to 1126.4M, while decreasing the computational cost from 934.5G to 858.2G FLOPs. This corresponds to a reduction of 76.3G FLOPs, or approximately 8.16%. Meanwhile, VulnDG improves the mIoU by 1.71, 2.31, and 1.26 points on Cityscapes, BDD100K, and Mapillary, respectively, increasing the average performance from 68.00% to 69.76%. These results indicate that the generalization improvements of VulnDG are not obtained by increasing the overall model capacity. Instead, the proposed vulnerability diagnosis, semantic calibration, and EMA-guided retention provide stronger robustness with a slightly smaller model size and lower computational cost.

Table 8: Efficiency and performance comparison with MFuser under the synthetic-to-real setting. Both methods use DINOv2-L as the frozen VFM and SIGLIP-L as the VLM. Params denotes the number of total parameters, and FLOPs are measured using a single 512×512 input during inference.

Method	Efficiency		**synthetic-to-real**			
	Params (M)↓	FLOPs (G)↓	$G \to C$	$G \to B$	$G \to M$	Avg.
MFuser	1127.5	934.5	71.10	61.19	71.71	68.00
VulnDG	**1126.4**	**858.2**	**72.81**	**63.50**	**72.97**	**69.76**

E More Qualitative Results

Confidence Heatmap Visualization. As shown in Figure 4, we overlay the predicted softmax probability of the ground-truth class on each input image, where blue and red indicate low and high confidence, respectively. MFuser and VulnDG use the same color scale within each row for a fair comparison. Compared with MFuser, VulnDG produces stronger and more spatially coherent responses over the target objects, with high-confidence regions better aligned with object boundaries and reduced activation in irrelevant background areas. Moreover, VulnDG maintains more complete responses over partially occluded and visually ambiguous object regions, indicating improved robustness to appearance and context variations.

Additional Qualitative Results on Cityscapes and BDD100K. Figure 5 presents additional qualitative comparisons under the $G \to C$ and $G \to B$ settings. Compared with tqdm and MFuser, VulnDG produces more spatially coherent predictions across diverse urban scenes, including daytime, low-illumination, and weather-degraded environments. In particular, our method better preserves the continuity of large semantic regions such as road, sidewalk, building, vegetation, and sky, while yielding cleaner boundaries around vehicles and roadside structures. The competing methods occasionally exhibit fragmented road regions, inaccurate sidewalk–road boundaries, or locally inconsistent predictions under strong appearance variations. By contrast, VulnDG more closely follows the ground-truth semantic layout, indicating that the diagnosed vulnerable relations are effectively calibrated without sacrificing the stability of already reliable regions.

Additional Qualitative Results on Mapillary. Figure 6 further compares the segmentation results under the $G \to M$ setting. Mapillary contains substantial variations in camera viewpoint, road structure, scene scale, illumination, and urban appearance, making it a challenging unseen domain. Across these diverse examples, VulnDG consistently recovers more complete road and sidewalk regions and produces more accurate semantic boundaries among buildings, vegetation, vehicles, and other foreground objects. It also reduces the local category confusion and spatial fragmentation observed in tqdm and MFuser, especially near object boundaries and in distant or small-scale regions. These results demonstrate that the proposed vulnerability-aware calibration improves not only the overall quantitative performance but also the structural consistency and visual reliability of cross-domain segmentation predictions.

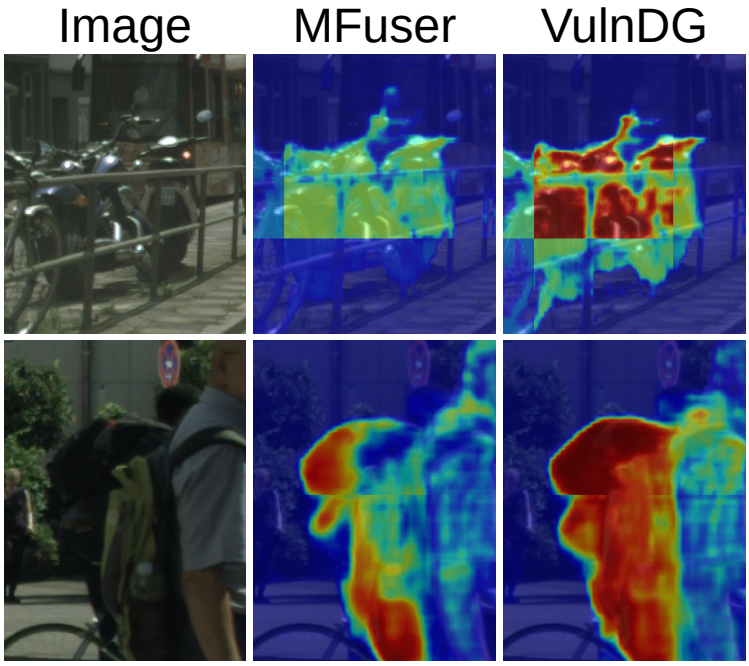

Figure 4: Qualitative comparison of per-pixel class-confidence heatmaps on two source-diagnosed vulnerable classes from the Cityscapes validation set: motorcycle (top) and rider (bottom).

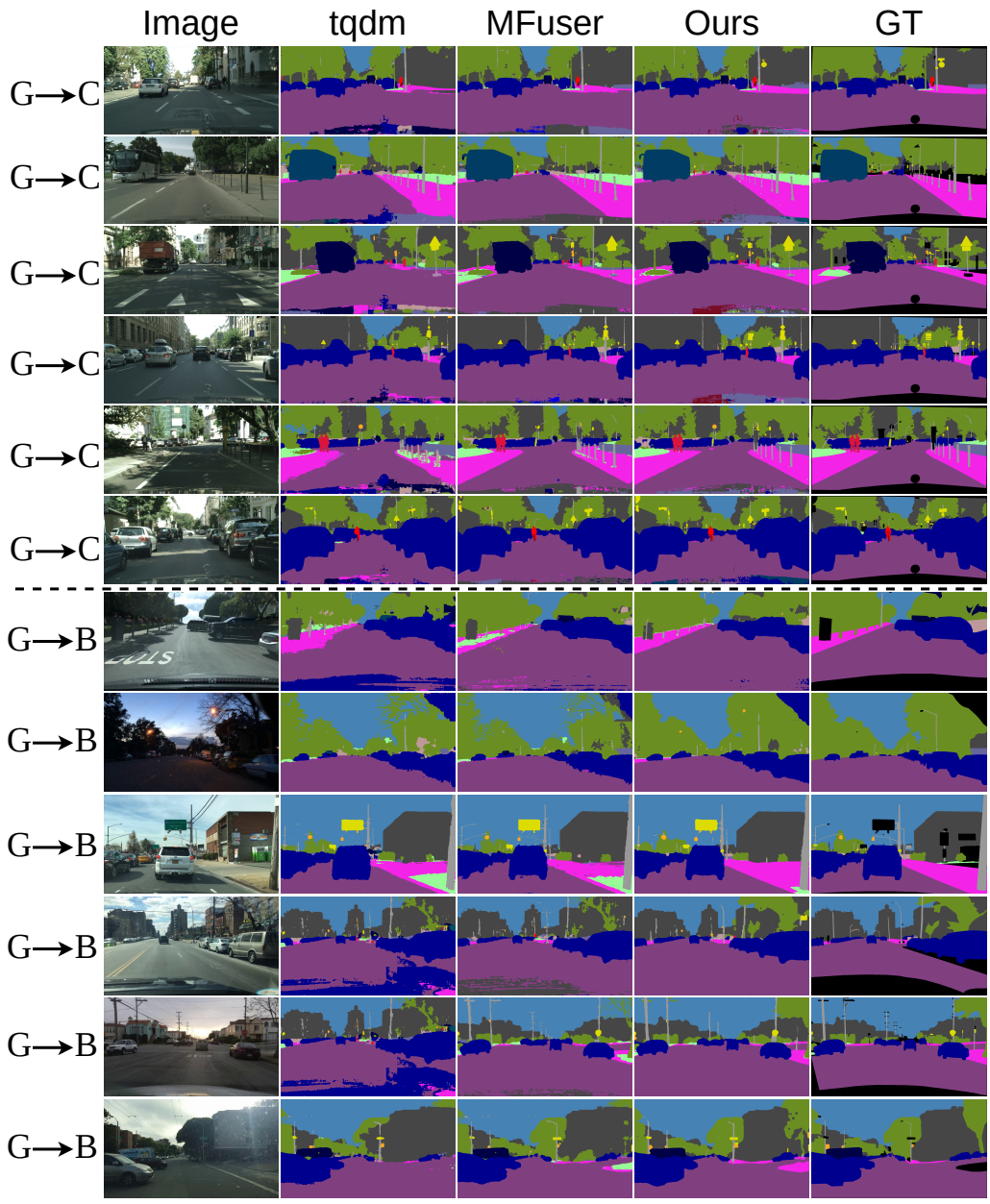

Figure 5: Qualitative results on unseen domains under the G → {B,C} setting. VulnDG is compared with tqdm and MFuser.

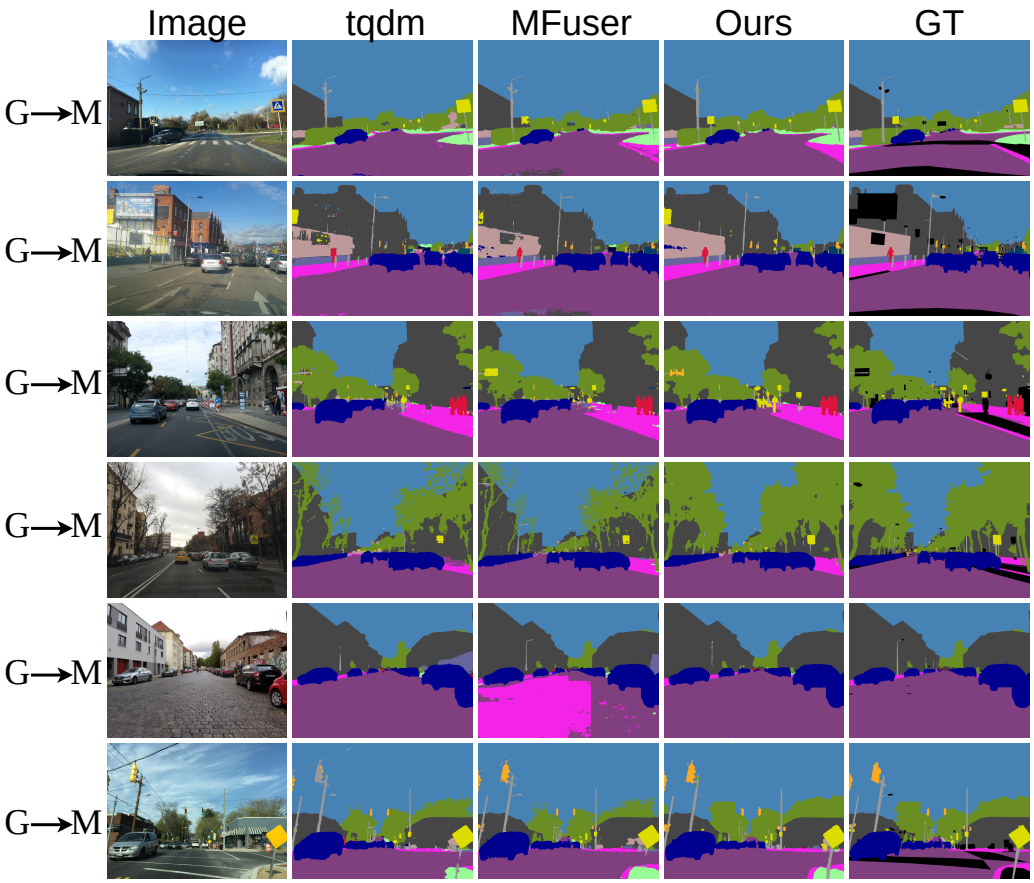

Figure 6: Qualitative results on unseen domains under the G $\rightarrow$ M setting. VulnDG is compared with tqdm and MFuser.