FuseAdapt: Parameter-Efficient Multimodal Fusion for Semantic Segmentation with Missing Modalities

Xin Zhang
National University of Singapore
x.zhang@u.nus.edu

Robby T. Tan
National University of Singapore
robby.tan@nus.edu.sg

Abstract

Multi-modal semantic segmentation benefits from complementary sensors, but this benefit is fragile when some modalities are degraded or missing at test time. Many existing methods rely on dedicated fusion modules or multi-branch cross-modal encoders, and the fusion often becomes unreliable when modalities are absent, leading to substantial performance drops. We propose **FuseAdapt**, a parameter-efficient framework that performs fusion within the adaptation space of a frozen vision foundation model (VFM), exploring whether VFM priors can serve as an effective shared backbone for multi-modal segmentation. Instead of attaching a standalone fusion network, FuseAdapt assigns each modality a lightweight adaptation branch and performs *adaptation-space fusion* by keeping Q/V low-rank adaptation private to each modality and routing only FFN low-rank residual activations across modalities inside the frozen backbone. The pretrained VFM remains fixed while modality-specific corrections and explicit cross-modal interaction are learned through the adaptation paths. To handle missing modalities, we further introduce **Predictive Modality Modeling (PMM)**, a direct representation-level objective for cross-modal redundancy. During training, PMM masks a target modality and predicts its embedding from the visible ones, supervised by a teacher target computed under full-modality input. PMM is used only during training and adds no inference-time overhead. Across three multi-modal segmentation benchmarks (MCubeS, DELIVER, MUSES), FuseAdapt consistently improves both modality-complete and modality-incomplete segmentation with a small trainable budget, yielding up to +3.01 mIoU under full modalities and +11.25 mIoU under missing modalities over prior state-of-the-art. Project page: https://devinxzhang.github.io/FuseAdapt-Page/.

1 Introduction

Scene understanding often benefits from combining multiple sensing modalities [Huang et al., 2022, Jia et al., 2023, Park et al., 2017, Zhang et al., 2018]. RGB [Wei et al., 2024, Zhang and Tan, 2025, Zhang et al., 2026, 2024a, Li et al., 2025a, Yuan et al., 2026a, Zhou et al., 2026, Zhang et al., 2024b, Zhang and Chen, 2023], depth [Cao et al., 2021, Song et al., 2022, Ying and Chuah, 2022], LiDAR [Yan et al., 2022, Wang et al., 2022, Li et al., 2023], and event streams [Xia et al., 2023, Zhang et al., 2021a, Li et al., 2025b] provide complementary cues about appearance, geometry, and motion. In practice, however, this multi-sensor benefit is fragile: a modality may be noisy [Yang et al., 2026a,b], degraded, spatially imperfect, or unavailable at inference time. Such cases are common in autonomous driving [Caesar et al., 2020, Li et al., 2026] and embodied settings [Wang et al., 2024], where illumination, weather [Du et al., 2026, Teng et al., 2025, Yang et al., 2025], texture, sensor failure, or calibration conditions [Yuan et al., 2025a] can affect sensors unequally. A practical multi-modal segmentation model should therefore work in two settings: modality-complete semantic segmentation (MCSS), where all modalities are available and should be fused effectively,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

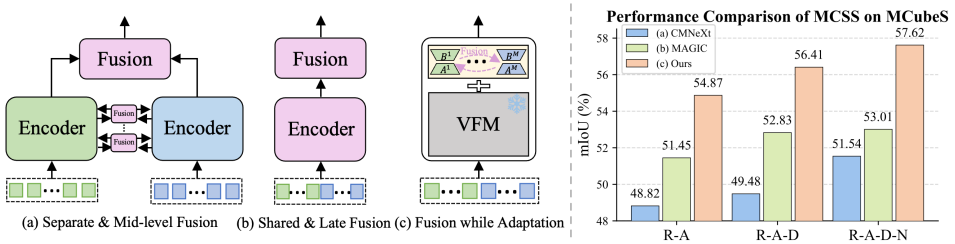


Figure 1: Comparison of multi-modal fusion strategies. (a) Separate encoders with mid-level fusion can replicate backbone computation across modalities. (b) Shared-encoder or late-fusion designs reduce this cost, but can make fusion depend strongly on the dominant or cleanest modality. (c) FuseAdapt keeps the VFM frozen and performs cross-modal exchange through trainable adaptation paths. Right: MCSS on MCubeS, where FuseAdapt improves mIoU across modality combinations.

and modality-incomplete semantic segmentation (MISS), where only a subset is available and the model should remain reliable.

Existing multi-modal segmentation methods often trade off modality-specific modeling, compactness, and robustness. Separate-encoder designs fuse intermediate features from different modalities [Broedermann et al., 2023, Zhang et al., 2023a], but can increase parameters and computation as the number of sensors grows. RGB-centric methods treat RGB as the main input and use other modalities as auxiliary cues [Zhang et al., 2023b], which can be limiting when RGB is degraded or absent. Shared-encoder and late-fusion methods improve compactness [Zheng et al., 2024c,b], but their shared stream can favor the most informative or easiest modality, making the contribution of other sensors less stable. For missing-modality robustness, recent methods go beyond modality dropout by using text anchors [Zheng et al., 2024b], subset consistency [Zheng et al., 2024c,a], or modality regularization [Zheng et al., 2025]. These strategies provide indirect supervision for dense prediction: text anchors are coarse, and subset consistency can be affected by cross-modal mismatch.

We study how to separate sensor-specific adaptation from cross-sensor communication within a shared frozen backbone. **FuseAdapt** keeps modality-specific Q/V low-rank residuals private and restricts receiver-conditioned, token-wise exchange to FFN low-rank residual activations (Fig. 1(c)). The frozen dense FFN transformation stays on the receiving stream, while routing is normalized over the available sensors. We call this constraint on the trainable communication path *adaptation-space fusion*. The exchanged tensors remain intermediate features; the contribution is the separation of private attention adaptation and cross-modal residual communication, rather than a new tensor domain or the individual use of LoRA and routing. A shared VFM supplies the pretrained representation, and controlled experiments with MiT backbones test whether the benefit depends on that representation.

To handle missing modalities, we introduce **Predictive Modality Modeling (PMM)** (Fig. 2(c)). PMM masks a target modality during training and predicts its representation from the visible modalities, supervised by a teacher target computed under full-modality input. PMM applies the joint-embedding predictive principle at modality level: target-specific queries predict an unobserved sensor's latent representation from visible context. Its EMA teacher supplies a temporally smoothed, stop-gradient target. After projection onto a common token grid, each query can attend to all visible positions rather than imposing same-index visible-to-target regression. This does not remove the common-grid requirement of routing and averaging. The teacher and predictor are used only during training.

We evaluate FuseAdapt on MCubeS [Liang et al., 2022], DELIVER [Zhang et al., 2023b], and MUSES [Brödermann et al., 2024] under both MCSS and MISS protocols. FuseAdapt yields up to +3.01 mIoU under full modalities and +11.25 mIoU under missing modalities over prior state-of-the-art, with ablations isolating the contributions of modality-specific adaptation, routed cross-modal exchange, PMM, and the choice of frozen VFM. Our contributions are:

- FuseAdapt separates modality-private Q/V adaptation from receiver-conditioned cross-modal exchange restricted to FFN low-rank residual activations, with routing adjusted to the visible sensor subset.

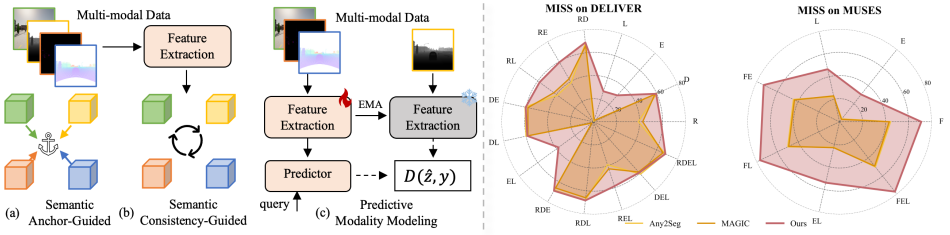


Figure 2: Comparison of designs for robustness under missing modalities (MISS). (a) Semantic-anchor methods align modality features to text embeddings, but text anchors can be coarse for pixel-level segmentation. (b) Semantic-consistency methods enforce agreement across different modality subsets, but such constraints can be affected by cross-modal mismatch. (c) Predictive Modality Modeling (PMM) casts MISS training as conditional representation prediction. Learnable query tokens aggregate cues from visible modalities to predict a missing-modality embedding, improving MISS performance on DELIVER and MUSES.

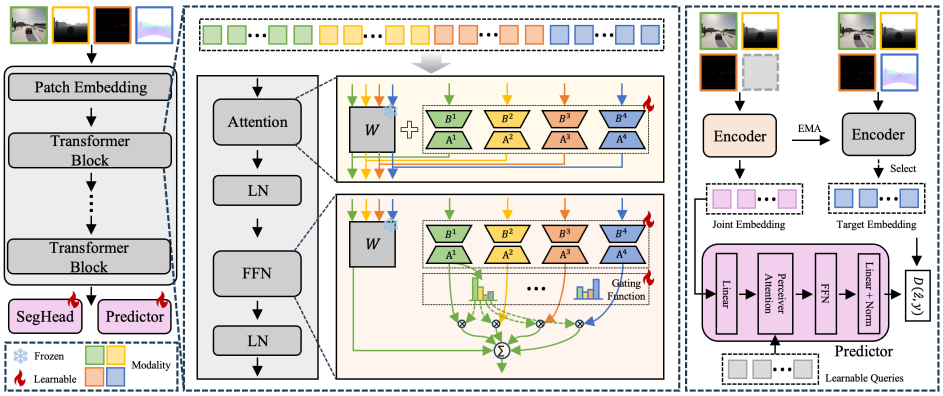
- PMM trains missing-modality robustness through target-conditioned latent prediction. Controls distinguish its prediction target from output distillation and generic fused-feature matching.
- Experiments on three benchmarks, same-backbone PEFT and information-flow controls support the design. We additionally characterize its inference cost and the scope of its robustness.

2 Related Work

Multi-Modal Semantic Segmentation Multi-modal semantic segmentation uses complementary RGB [Wei et al., 2024, Zhang and Tan, 2025, Qian et al., 2024, Zhang et al., 2022, Deng et al., 2020], depth [Wang et al., 2020, Ying and Chuah, 2022, Song et al., 2022, Cao et al., 2021, Park et al., 2017], LiDAR [Li et al., 2022, 2023, Zhuang et al., 2021, Yan et al., 2022, Wang et al., 2022], thermal [Hui et al., 2023, Zhang et al., 2023c, 2021b, Wu et al., 2022], and event cues [Xia et al., 2023, Zhang et al., 2021a, Jia et al., 2023]. Prior work often uses fixed modality pairs such as RGB-depth [Ying and Chuah, 2022, Song et al., 2022], RGB-centric fusion [Zhang et al., 2023b], or shared-backbone fusion with auxiliary alignment [Zheng et al., 2024b,c,a]. In contrast, FuseAdapt keeps a frozen VFM and performs cross-modal exchange through trainable adaptation paths, avoiding a standalone fusion network.

Missing or Degraded Modalities Existing approaches use modality dropout, cross-modal distillation, and consistency regularization [Liu et al., 2024, Wang et al., 2023, Zheng et al., 2024b,c,a]. CRM [Shin et al., 2024] combines complementary RGB-thermal masking with clean-to-masked self-distillation; M3L [Maheshwari et al., 2024] uses a full-modality teacher to supervise masked-modality segmentation. PMM instead conditions on a designated missing sensor and predicts its full-context latent representation. This follows the latent-context prediction principle of I-JEPA [Assran et al., 2023], applied to modalities rather than image regions. Masking, EMA, and teacher–student optimization are established mechanisms; the prediction target and modality conditioning distinguish our objective.

Parameter-Efficient Adaptation and Fusion Prompt tuning, adapters, and LoRA reduce the cost of adapting pretrained models [Lester et al., 2021, Jia et al., 2022, Houlsby et al., 2019, Chen et al., 2022, Hu et al., 2022, Yan et al., 2025]. StitchFusion [Li et al., 2024] transfers dense features between encoder streams through MultiAdapters. MLE-SAM [Zhu et al., 2024] combines modality-specific LoRA experts and routing with multi-scale feature/head fusion. MoLE [Wu et al., 2024] composes task/concept LoRAs within a single stream. Concurrent M4-SAM [Liu et al., 2026] combines modality-dispatched convolutional experts with gated feature fusion for RGB-D video, while MILE [Muralidhara et al., 2026] selects isolated continual-task experts. FuseAdapt instead keeps Q/V residuals private and restricts directed source-to-receiver communication to FFN low-rank residuals among currently visible sensors. Section 4.4 tests this constraint against routing all low-rank paths or complete FFN features.



(a) FuseAdapt (b) PMM

Figure 3: **Overview of our FuseAdapt framework.** (a) *Adaptation-Space Fusion:* Multi-modal inputs (e.g., RGB, depth, event, LiDAR) share a frozen VFM backbone. Each modality attaches a lightweight adaptation branch inside the transformer blocks, and a receiver-conditioned router mixes only FFN low-rank residual activations. Fusion is thus performed inside the adaptation path while keeping the backbone frozen, and the fused features are sent to a shared segmentation head. For clarity, we show the routing for one modality; the others follow the same design. (b) *PMM:* To handle missing modalities, an EMA teacher provides the target embedding for a selected modality. A student and a Perceiver-style predictor use the visible modalities to predict this embedding via learnable queries. PMM is trained with a distance loss $D(\hat{z}, z)$.

3 Proposed Method

We address robust multi-modal semantic segmentation in two regimes: modality-complete semantic segmentation (MCSS), where all modalities are available, and modality-incomplete semantic segmentation (MISS), where only a subset is available at test time. Our goal is to adapt a frozen vision foundation model (VFM) into a multi-modal learner that can fuse available modalities with a small trainable budget and learn redundancy for missing-modality robustness.

As shown in Fig. 3, the input contains four aligned image-like modalities of the same scene: RGB $\mathbf{R} \in \mathbb{R}^{h\times w\times 3}$, depth $\mathbf{D} \in \mathbb{R}^{h\times w\times C^D}$, LiDAR projection $\mathbf{L} \in \mathbb{R}^{h\times w\times C^L}$, and event stack $\mathbf{E} \in \mathbb{R}^{h\times w\times C^E}$, with pixel labels $\mathbf{Y}$ over K classes. Following [Zhang et al., 2023b], we set $C^D = C^L = C^E = 3$ and feed all modalities as image-like inputs to a shared frozen VFM.

FuseAdapt has two components. First, an *adaptation-space fusion design* (§3.1) performs fusion in the adaptation space of the frozen VFM. Each modality has lightweight trainable adaptation paths, and cross-modal exchange is implemented by routing FFN low-rank residual activations across modalities inside Transformer blocks. The pretrained backbone remains fixed. Second, *Predictive Modality Modeling (PMM)* improves MISS by training the model to predict a masked modality representation from visible modalities, using a teacher target computed with full-modality context. PMM is used only during training and discarded at inference.

3.1 Adaptation-Space Fusion

We realize adaptation-space fusion through modality-specific low-rank adaptation paths inside the frozen VFM, with cross-modal exchange implemented by routed FFN low-rank residual activations. We keep self-attention adaptation modality-specific to preserve sensor-specific correction, and place cross-modal exchange in the FFN adaptation path.

Self-Attention Adaptation Let $\mathcal{M} = \{r, d, e, l\}$ denote the modality set. At a Transformer block, let $\mathbf{X}^m \in \mathbb{R}^{N\times d}$ be the token features of modality m, where all modalities have the same projected token count N at this block and d is the hidden dimension. Routing and averaging require this common grid; arbitrary native token lengths are not supported. We omit the block index and bias terms for clarity. For each modality, we add trainable low-rank updates only to the query and value

projections:

$$\mathbf{Q}^m = \mathbf{X}^m(\mathbf{W}_Q + \Delta\mathbf{W}_Q^m), \quad \mathbf{V}^m = \mathbf{X}^m(\mathbf{W}_V + \Delta\mathbf{W}_V^m), \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ are frozen pretrained weights. The updates are parameterized as

$$\Delta\mathbf{W}_Q^m = \mathbf{A}_Q^m\mathbf{B}_Q^m, \quad \Delta\mathbf{W}_V^m = \mathbf{A}_V^m\mathbf{B}_V^m, \tag{2}$$

where $\mathbf{A}_Q^m, \mathbf{A}_V^m \in \mathbb{R}^{d\times r}$, $\mathbf{B}_Q^m, \mathbf{B}_V^m \in \mathbb{R}^{r\times d}$, and $r \ll d$. The key and output projections, $\mathbf{W}_K$ and $\mathbf{W}_O$, remain frozen and shared. The attention output is

$$\mathrm{MHSA}^m(\mathbf{X}^m) = \mathrm{softmax}\left(\frac{\mathbf{Q}^m(\mathbf{X}^m\mathbf{W}_K)^\top}{\sqrt{d}}\right)\mathbf{V}^m\mathbf{W}_O. \tag{3}$$

This keeps attention adaptation separated by modality while preserving the pretrained attention structure.

Feedforward Routing Fusion Cross-modal exchange is introduced in the FFN adaptation path. Let $\mathbf{W}_1 \in \mathbb{R}^{d\times l}$ and $\mathbf{W}_2 \in \mathbb{R}^{l\times d}$ be the frozen FFN weights. Each modality has trainable low-rank FFN updates:

$$\Delta\mathbf{W}_1^m = \mathbf{A}_1^m\mathbf{B}_1^m, \quad \Delta\mathbf{W}_2^m = \mathbf{A}_2^m\mathbf{B}_2^m, \tag{4}$$

where $\mathbf{A}_1^m \in \mathbb{R}^{d\times r}$, $\mathbf{B}_1^m \in \mathbb{R}^{r\times l}$, $\mathbf{A}_2^m \in \mathbb{R}^{l\times r}$, $\mathbf{B}_2^m \in \mathbb{R}^{r\times d}$, and $r \ll \min(d, l)$.

Let $\mathcal{V} \subseteq \mathcal{M}$ be the nonempty available subset. For receiving modality $m \in \mathcal{V}$ and token i, router logits $g^m(\mathbf{X}^m)_i$ define

$$\alpha_{ij}^m = \frac{\exp(g^m(\mathbf{X}^m)_{ij})}{\sum_{k\in\mathcal{V}}\exp(g^m(\mathbf{X}^m)_{ik})}, \quad j \in \mathcal{V}, \qquad \alpha_{ij}^m = 0 \quad (j \notin \mathcal{V}). \tag{5}$$

Thus unavailable sources are masked before normalization. Omitting block indices, biases, and dropout, the FFN becomes

$$\mathbf{H}^m = \sigma\left(\mathbf{X}^m\mathbf{W}_1 + \sum_{j\in\mathcal{V}}\boldsymbol{\alpha}_j^m \odot (\mathbf{X}^j\Delta\mathbf{W}_1^j)\right), \tag{6}$$

$$\mathbf{F}^m = \mathbf{H}^m\mathbf{W}_2 + \sum_{j\in\mathcal{V}}\boldsymbol{\alpha}_j^m \odot (\mathbf{H}^j\Delta\mathbf{W}_2^j), \tag{7}$$

where σ is the FFN activation and $\odot$ broadcasts token weights over channels. The routed objects are residual activations, not parameter matrices or complete FFN features. Dense transformations operate on the receiving stream; their weights remain frozen even though stream activations evolve through successive blocks. After the final block, we symmetrically aggregate visible outputs,

$$\hat{\mathbf{X}}_\mathcal{V} = \frac{1}{|\mathcal{V}|}\sum_{m\in\mathcal{V}}\hat{\mathbf{X}}^m, \tag{8}$$

and pass them to the shared segmentation decoder. The same communication rule applies to full and partial inputs.

3.2 Predictive Modality Modeling

While adaptation-space fusion handles full-modality inputs, MISS also requires the model to behave reliably when some modalities are absent. PMM addresses this by adding a representation-level prediction objective during training. Given visible modalities, the model predicts the representation of a masked target modality, supervised by a teacher target computed with full-modality context. This encourages cross-modal redundancy beyond segmentation supervision alone.

After the full-input warm-up, training mixes full and masked inputs. In the masked branch, modalities are independently retained with probability 0.5, conditioned on a nonempty visible subset $\mathcal{V}$. When $\mathcal{V} \neq \mathcal{M}$, we sample a target $t \in \mathcal{M} \setminus \mathcal{V}$ uniformly and aggregate visible features $\hat{\mathbf{X}}^m \in \mathbb{R}^{N\times C}$ as above. When all modalities are visible, segmentation supervision remains active and no missing-target loss is computed. Appendix F.3 gives the masking procedure.

The teacher target $y_t \in \mathbb{R}^{N\times C}$ for the masked modality t is produced by an EMA teacher that takes the *full-modality* input. The teacher shares the frozen VFM with the student (no duplication of

Table 1: Results in MCSS setting on MCubeS. The best results are highlighted in **bold**. (R: RGB; A: AoLP; D: DoLP; N: NIR.)

Method	Backbone	#Trainable Params (M)	Modals			Mean
			RA	RAD	RADN	
MCubeSNet [Liang et al., 2022]	MiT-B4	65.65	39.10	42.00	42.86	41.32
CWSAM [Pu et al., 2025]	ViT-B	15.29	49.78	48.27	50.59	49.55
MLE-SAM [Zhu et al., 2024]	Hiera-B+	20.79	50.61	50.89	51.02	50.84
CMNeXt [Zhang et al., 2023b]	MiT-B2	58.73	48.82	49.48	51.54	49.95
Any2Seg [Zheng et al., 2024b]	MiT-B2	24.73	48.95	50.67	51.91	50.51
MemorySAM [Liao et al., 2025]	Hiera-B+	14.50	51.20	52.20	52.88	52.09
MAGIC [Zheng et al., 2024c]	MiT-B2	24.73	51.45	52.83	53.01	52.43
StitchFusion [Li et al., 2024]	MiT-B4	68.02	52.68	53.26	53.92	53.29
Ours (wo FuseAdapt)	ViT-B	9.57	49.15	49.98	51.16	50.10
Ours (shared FuseAdapt)	ViT-B	10.64	54.87	54.85	54.69	54.80
Ours (separate FuseAdapt)	ViT-B	13.86	54.87	55.26	56.09	55.41
Ours	ViT-B	14.01	**54.87**	**56.41**	**57.62**	**56.30**

backbone weights) and maintains an EMA copy only of the adaptation branches. Letting θ_s denote the student FuseAdapt parameters and θ_τ those of the teacher, the EMA update is

$$\theta_\tau \leftarrow \mu\,\theta_\tau + (1-\mu)\,\theta_s, \qquad \mu = 0.999, \tag{9}$$

applied at every iteration. The target is stop-gradient. Full context is a useful training prior on average, not a guarantee of superior predictions for every sample, especially when a present sensor is corrupted (Appendix C.1).

Query-Based Prediction A per-token linear regressor imposes same-index visible-to-target correspondence on the common grid. We instead use modality-specific learnable queries $Q_t \in \mathbb{R}^{N\times C}$, each of which attends globally to $\hat{\mathbf{X}}_\mathcal{V}$ through a Perceiver-style predictor:

$$\hat{z}_t = g_\phi(Q_t, \hat{\mathbf{X}}_\mathcal{V}), \qquad \hat{z}_t \in \mathbb{R}^{N\times C}. \tag{10}$$

This permits retrieval from any visible position while retaining a common target grid. It does not natively align different fields of view or temporal offsets. The PMM loss is a token-wise cosine distance:

$$\mathcal{L}_{\text{pmm}} = 1 - \frac{1}{N}\sum_{i=1}^{N} D\big(\hat{z}_t^{(i)}, \text{sg}[y_t^{(i)}]\big), \tag{11}$$

where $D(\cdot,\cdot)$ denotes cosine similarity. The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{pmm}}. \tag{12}$$

At inference time, PMM is removed and the model uses the available modalities directly.

4 Experiments

Datasets We evaluate on three multi-modal segmentation benchmarks covering both modality-complete semantic segmentation (MCSS) and modality-incomplete semantic segmentation (MISS). DELIVER [Zhang et al., 2023b] contains synchronized RGB, depth, LiDAR, and event inputs for 25-class urban-scene segmentation, with 3,983/2,005/1,897 train/val/test samples and diverse adverse conditions such as fog, rain, night, blur, exposure changes, and LiDAR jitter [Yuan et al., 2026b]. MCubeS [Liang et al., 2022] contains 20-class material segmentation with RGB, NIR, DoLP, and AoLP modalities [Lin et al., 2025], using 302/96/102 train/val/test samples. MUSES [Brödermann et al., 2024] provides synchronized frame-camera, event-camera, and MEMS-LiDAR data with 2D panoptic annotations, enabling evaluation with heterogeneous sensors.

Implementation Details We use AdamW and train on A5000 GPUs: 8 GPUs for DELIVER and 4 GPUs for MCubeS/MUSES. Batch size is 1 per GPU for DELIVER and 4 for MCubeS/MUSES. The initial learning rate is 4×10^{-4} for DELIVER/MUSES and 3.5×10^{-4} for MCubeS, with a 10-epoch warm-up at $0.1\times$. We train for 200 epochs on DELIVER/MUSES and 500 epochs on MCubeS. Input resolutions are 1024×1024 for DELIVER/MUSES and 512×512 for MCubeS. We use standard augmentations, including random resizing (0.5–2.0), horizontal flipping, color jitter, blur, and random cropping. FuseAdapt uses rank 8 for DELIVER and MCubeS, and 4 for MUSES. We use DINOv2-B for DELIVER/MCubeS and DINOv2-S for MUSES [Oquab et al., 2023] as frozen VFM backbones.

Table 2: Results in MCSS setting on MUSES. The best results are highlighted in **bold**. (F: frame camera, E: event camera, L: LiDAR.)

Method	Backbone	#Trainable Params (M)	Modals				Mean
			F	FE	FL	FEL	
CWSAM [Pu et al., 2025]	ViT-B	15.29	55.41	41.77	40.69	49.98	46.96
CMNeXt [Zhang et al., 2023b]	MiT-B0	10.30	43.37	43.39	47.03	46.66	45.11
MLE-SAM [Zhu et al., 2024]	Hiera-B+	20.79	73.95	74.73	75.42	74.80	74.73
Ours (wo FuseAdapt)	ViT-S	2.42	62.46	63.00	64.72	65.14	63.83
Ours (shared FuseAdapt)	ViT-S	2.69	74.99	74.96	76.23	76.59	75.69
Ours (separate FuseAdapt)	ViT-S	3.22	74.99	74.81	76.13	76.38	75.58
Ours	ViT-S	3.29	**74.99**	**75.36**	**76.94**	**77.28**	**76.14**

Table 3: Results in MCSS setting on DELIVER. The best results are highlighted in **bold**. (R: RGB; D: Depth; E: Event; L: LiDAR.)

Method	Backbone	#Trainable Params (M)	Modals						Mean
			RD	RE	RL	RDE	RDL	RDEL	
HRFuser [Broedermann et al., 2023]	HRFuser-T	31.61	51.88	42.22	43.13	51.83	52.72	52.97	49.13
CWSAM [Pu et al., 2025]	ViT-B	15.29	58.64	-	-	56.22	-	55.43	-
MLE-SAM [Zhu et al., 2024]	Hiera-B+	20.79	63.57	-	-	62.69	-	64.08	-
MemorySAM [Liao et al., 2025]	Hiera-B+	14.50	63.48	-	-	62.42	-	65.38	-
TokenFusion [Wang et al., 2022]	MiT-B2	26.01	60.25	45.63	53.01	-	-	-	-
CMX [Zhang et al., 2023a]	MiT-B2	66.57	62.67	56.62	56.37	-	-	-	-
CMNeXt [Zhang et al., 2023b]	MiT-B2	58.73	63.58	57.48	58.04	64.44	65.50	66.30	-
Zheng et al. [Zheng et al., 2025]	MiT-B2	24.73	66.46	56.06	-	-	-	-	-
MAGIC [Zheng et al., 2024c]	MiT-B2	24.73	66.89	58.48	57.75	66.24	**67.63**	67.66	-
StitchFusion [Li et al., 2024]	MiT-B2	26.50	65.75	57.44	58.03	66.03	66.65	68.18	63.68
Any2Seg [Zheng et al., 2024b]	MiT-B2	24.73	67.12	57.83	57.92	66.75	67.20	68.25	64.18
OmniSegmentor [Yin et al., 2025]	DFormer-L	39.00	64.70	58.60	59.20	65.90	67.20	68.00	63.93
Ours	ViT-B	14.01	**67.33**	**59.62**	**60.11**	**67.16**	67.29	**68.33**	**64.97**

Evaluation Protocol We report two settings. In MCSS, all modalities are available at training and inference, measuring full-modality fusion. In MISS, the model is trained with all modalities and evaluated on modality subsets at inference, measuring robustness when sensors are unavailable.

Comparison Context Cross-backbone headline comparisons reflect both the method and the pretrained representation. We separately report internally trained controls with frozen MiT-B0/B2 and a common SegFormer decoder [Xie et al., 2021], recipe, and checkpoint rule (Section 4.4). DINOv2-S was chosen on MUSES to keep the trainable budget comparable to MiT-B0 baselines, not to match backbone capacity. A matched DINOv2-B run reaches 80.16 versus 77.28 with DINOv2-S (Appendix A).

4.1 Modality-Complete Semantic Segmentation (MCSS) Results

We first evaluate MCSS on MCubeS, MUSES, and DELIVER, where all modalities are available during training and inference. As shown in Tabs. 1, 2, and 3, FuseAdapt achieves the best reported mean performance among the compared methods on all three benchmarks with a small trainable budget. On MCubeS, FuseAdapt obtains 57.62 mIoU with all four modalities (RADN) and 56.30 mean mIoU, compared with 53.29 for StitchFusion and 52.43 for MAGIC. On DELIVER, it achieves 64.97 mean mIoU, outperforming StitchFusion (63.68) and Any2Seg (64.18). On MUSES, it reaches 76.14 mean mIoU, improving over the prior reported results across all modality combinations. These results indicate that adapting a frozen VFM with modality-specific paths and routed adaptation exchange provides effective full-modality fusion.

4.2 Modality-Incomplete Semantic Segmentation (MISS) Results

We next evaluate MISS, where models are trained with all modalities and tested on arbitrary modality subsets. As shown in Tabs. 5, 4, and 6, FuseAdapt improves performance across a broad range of missing-modality settings. On MCubeS, FuseAdapt achieves 52.89 mean mIoU, outperforming

Table 4: Results in MISS setting on MUSES. All methods are trained with three modalities. The best results are highlighted in **bold**.

Method	Backbone	#Trainable Params (M)	Modals							Mean
			F	E	L	FE	FL	EL	FEL	
CMX [Zhang et al., 2023a]	MiT-B0	10.32	2.52	2.35	3.01	41.15	41.25	2.56	42.27	19.30
CMNeXt [Zhang et al., 2023b]	MiT-B0	10.30	3.50	2.77	2.64	6.63	10.28	3.14	46.66	10.80
MAGIC [Zheng et al., 2024c]	MiT-B0	3.72	43.22	2.68	22.95	43.51	49.05	22.98	49.02	33.34
Any2Seg [Zheng et al., 2024b]	MiT-B0	3.72	44.40	3.17	22.33	44.51	49.96	22.63	50.00	33.86
MLE-SAM [Zhu et al., 2024]	Hiera-B+	20.79	69.67	1.50	5.55	69.50	74.11	5.55	74.80	42.95
Zheng et al. [Zheng et al., 2025]	MiT-B0	3.72	60.33	**33.15**	42.59	47.19	53.47	39.59	47.89	46.32
BiXFormer [Chen et al., 2025a]	MiT-B0	17.70	65.76	16.88	39.63	58.42	59.77	49.89	65.81	50.88
Ours	ViT-S	3.29	**71.09**	30.16	**46.59**	**73.24**	**76.72**	**54.04**	**77.28**	**61.30**

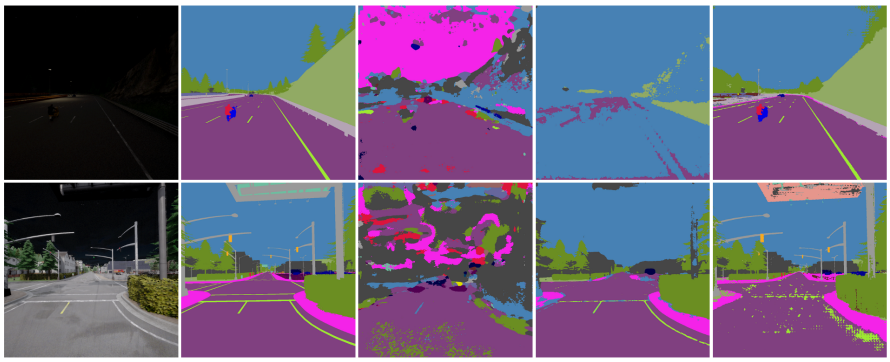


Figure 4: Qualitative results on DELIVER in MISS setting. The proposed method is compared with CMNeXt [Zhang et al., 2023b] and StitchFusion [Li et al., 2024]. The model is trained on four modalities $\{r, d, e, l\}$ and tested on $\{r\}$ (Row 1) and $\{r, l\}$ (Row 2).

MAGIC (38.00) and Zheng et al. (41.64). On DELIVER, it obtains 56.22 mean mIoU, compared with 45.04 for Any2Seg, 44.85 for MAGIC++, and 41.98 for StitchFusion. On MUSES, it reaches 61.30 mean mIoU with 3.29M trainable parameters, outperforming the reported baselines. The gains across different modality subsets suggest that the learned representation remains useful under changing sensor availability. Fig. 4 shows qualitative MISS results on DELIVER. With only RGB (Row 1) or RGB+LiDAR (Row 2), FuseAdapt produces cleaner predictions and more coherent scene layouts than CMNeXt and StitchFusion, especially in challenging night or fog cases where baselines show larger region-level errors.

4.3 Ablation Studies

We ablate FuseAdapt and PMM under the same training recipe and also examine the effect of changing the frozen VFM backbone.

FuseAdapt Variants (MCSS) Tabs. 1 and 2 compare variants within the same frozen-VFM framework. Without FuseAdapt, the model uses naive aggregation of per-modality features and performs worse. Shared FuseAdapt improves performance but uses one adaptation path for all modalities, limiting modality-specific correction. Separate FuseAdapt uses modality-specific paths without routed exchange; it improves over Shared on MCubeS, while their MUSES DINOv2-S means are similar. Full FuseAdapt performs best in both tables. This supports the contribution of both modality-specific adaptation and routed adaptation exchange.

PMM Variants (MISS) Tab. 5 evaluates PMM variants under MISS on MCubeS. Removing the predictor or replacing it with a linear predictor reduces mean mIoU from 52.89 to 43.67 and 49.71, respectively. This shows that the representation-level prediction objective is important, and that the query-based predictor is more effective than a simple linear prediction head. Additional ablations, including FuseAdapt rank and scalability, are provided in the supplementary material.

Additional VFMs To examine sensitivity to the frozen VFM choice, we keep the full pipeline fixed and replace DINOv2-B with EVA-CLIP-B [Sun et al., 2023] and NeCo-B [Pariza et al., 2024] on

Table 5: Results in MISS setting on MCubeS. All methods are trained with four modalities. The best results are highlighted in **bold**.

Method	Backbone	Trainable Params*	Modals															Mean
			R	A	D	N	RA	RD	RN	AD	AN	DN	RAD	RAN	RDN	ADN	RADN	
CMNeXt [Zhang et al., 2023b]	MiT-B2	58.73	1.86	1.54	2.51	2.28	47.96	43.67	45.90	6.99	7.58	9.95	50.07	48.77	48.83	8.06	51.54	25.17
MAGIC [Zheng et al., 2024c]	MiT-B2	24.73	51.91	0.32	34.52	2.66	52.24	52.16	52.57	1.98	36.01	37.09	52.48	52.82	52.73	37.57	53.01	38.00
Zheng et al. [Zheng et al., 2025]	MiT-B2	24.73	46.21	**37.66**	40.49	40.56	43.93	43.11	43.73	40.10	39.19	40.72	41.78	41.53	42.21	40.63	42.73	41.64
Ours (wo pred.)	ViT-B	14.01	51.02	13.60	22.47	33.61	53.16	54.92	54.04	26.37	38.26	40.24	55.33	55.67	56.61	43.02	56.66	43.67
Ours (linear pred.)	ViT-B	14.01	51.38	31.08	42.02	44.97	53.64	54.64	54.82	43.36	47.38	49.62	55.01	55.31	56.06	50.15	56.20	49.71
Ours	ViT-B	14.01	**54.96**	37.60	**47.14**	**49.12**	**56.17**	**57.23**	**56.29**	**47.96**	**51.65**	**52.74**	**57.19**	**56.80**	**57.58**	**53.30**	**57.62**	**52.89**

Table 6: Results in MISS setting on DELIVER. All methods are trained with four modalities. The best results are highlighted in **bold**.

Method	Backbone	#Trainable Params (M)	Modals															Mean
			R	D	E	L	RD	RE	RL	DE	DL	EL	RDE	RDL	REL	DEL	RDEL	
CMNeXt [Zhang et al., 2023b]	MiT-B2	58.73	22.50	50.59	3.16	2.86	66.33	22.92	22.50	50.80	50.83	3.15	66.27	66.38	22.92	50.98	66.33	37.90
GeminiFusion [Jia et al., 2024]	MiT-B2	42.18	15.89	54.73	1.70	1.70	66.93	16.24	15.80	54.83	54.76	1.70	66.92	66.93	16.18	54.86	66.92	37.07
MAGIC [Zheng et al., 2024c]	MiT-B2	24.73	42.72	58.39	1.90	1.62	66.10	42.79	42.72	58.44	58.39	1.90	66.11	66.10	42.80	58.44	66.10	44.97
MAGIC++ [Zheng et al., 2024a]	MiT-B2	27.51	41.10	58.12	2.14	1.64	67.33	41.13	41.13	58.32	58.12	2.15	67.35	67.33	41.17	58.31	67.34	44.85
Any2Seg [Zheng et al., 2024b]	MiT-B2	24.73	39.02	**60.11**	2.07	0.31	68.21	39.11	39.04	**60.92**	**60.15**	1.99	68.24	68.22	39.06	**60.95**	68.25	45.04
StitchFusion [Li et al., 2024]	MiT-B2	26.50	30.93	55.44	1.87	1.59	68.22	31.03	33.55	55.76	55.41	1.87	68.23	68.21	33.66	55.71	68.20	41.98
Zheng et al. [Zheng et al., 2025]	MiT-B2	24.73	55.56	57.74	**32.40**	**34.26**	59.69	47.70	47.78	48.99	50.26	34.37	53.42	54.12	45.24	47.33	51.65	48.03
Ours	ViT-B	14.01	**58.52**	59.48	30.73	27.67	**68.93**	**59.43**	**58.08**	60.39	58.89	**37.72**	**68.90**	**68.42**	**58.58**	59.22	**68.33**	**56.22**

MCubeS under MCSS. FuseAdapt achieves 56.02 mIoU with EVA-CLIP-B and 56.28 with NeCo-B, compared with 57.62 using DINOv2-B. Although absolute performance depends on the chosen VFM, the method remains competitive across different frozen backbones, reducing dependence on a single pretrained model.

4.4 Controlled Comparisons

Backbone and recipe controls. On MUSES with frozen MiT-B0, FuseAdapt obtains 57.54 mIoU versus 49.47 for Deep Prompt, 53.56 for Shared LoRA, and 54.80 for Adapter. These internal controls use the same decoder, split, augmentation, optimizer, schedule, PMM, and checkpoint rule, with 2.95/2.81/2.95/2.94M trainable parameters, respectively. External MiT-B0 references MAGIC and Any2Seg obtain 49.02/50.00. With MiT-B2, FuseAdapt obtains 70.60, compared with 58.28/55.36/53.40/59.36 for CMNeXt/MAGIC/StitchFusion/CHARM. External results share the backbone family but are not all recipe-matched; Appendix A gives definitions and protocols.

Where should information cross modalities? Dense-feature routing keeps Q/V private but mixes complete FFN features; All-path routed LoRA routes Q/V and FFN residuals. Both retain the backbone and decoder of the corresponding setting. FuseAdapt outperforms these alternatives on MCubeS and MUSES (Table 7), supporting the separation of private attention adaptation and residual-only exchange. These are internal mechanism controls, not reproductions of complete prior systems.

Prediction versus generic distillation. With the same backbone, masking, EMA teacher, segmentation supervision, and schedule, output-logit KD and fused-feature KD obtain 49.84 and 50.57 MISS mean mIoU on MCubeS, versus 52.89 for PMM. PMM improves output KD in all 15 nonempty subsets and fused-feature KD in 14/15. Removing EMA from the query predictor gives 50.59; the original per-token regressor gives 49.71. These controls support the target-conditioned prediction objective beyond generic full-to-partial matching. Table 8 summarizes the objectives; Appendix C gives every subset.

4.5 Robustness and Deployment Cost

Scope of robustness. On DELIVER's ten predefined weather/degradation conditions, FuseAdapt obtains 67.36 mean mIoU versus 67.19 for Any2Seg, leading in eight conditions but trailing at night and under-exposure. In a separate MCubeS latent-prediction test, shifting the visible sensors together by 16 pixels relative to the masked target reduces cosine similarity by 0.00051 for the query predictor versus 0.01636 for per-token regression. The linear model has higher clean cosine; the evidence

Table 7: Full-modality mIoU for information-flow variants (seed 3407).

Variant	MCubeS / DINOv2-B	MUSES / MiT-B0
Shared	54.69	53.56
Separate	56.09	54.31
Dense-feature routing	55.08	55.94
All-path routed LoRA	55.74	56.17
FuseAdapt	**57.62**	**57.54**

Table 8: MCubeS MISS mean mIoU over all 15 nonempty subsets. Only the auxiliary objective/target differs; full results are in Table 13.

Objective	Mean	Objective	Mean
Dropout only	43.67	Query without EMA	50.59
Per-token regression	49.71	Global-feature KD	50.57
Output KD	49.84	Full PMM	**52.89**

concerns lower shift sensitivity, not universally better reconstruction or arbitrary sensor misalignment. Appendix D provides both protocols and full tables.

End-to-end efficiency. On an idle H100, at 1024^2, batch 1, FP32, with 20 warm-ups and 100 timed runs, three-modality MiT-B0 FuseAdapt takes 21.05 ms versus 19.63 for CMNeXt and 26.69 for StitchFusion. It retains 3.91M unique deployed parameters and removes both PMM components. Parameter efficiency does not imply lowest latency or memory: Any2Seg and MAGIC are faster in this profile. Appendix E reports all measured costs and identifies reconstructed inference graphs.

5 Discussion and Conclusion

FuseAdapt separates modality-private attention adaptation from cross-modal exchange through FFN low-rank residuals. PMM complements this design with target-conditioned missing-modality latent prediction. Same-backbone controls support both choices, with improvements across the evaluated MiT scales.

Limitations and deployment scope. Absolute accuracy depends on the frozen representation and its biases; cross-backbone comparisons alone cannot isolate our contribution. Inference cost still grows with the number of sensor streams. Routing and averaging require a common projected grid, and the shift experiment measures latent prediction sensitivity, not robustness to arbitrary calibration, field-of-view, or temporal errors. DELIVER covers predefined adverse conditions, but neither those results nor a full-context EMA target guarantees safe behavior under correlated or out-of-distribution sensor failures, including device-level ones [Yuan et al., 2025b,c]. Deployment in robotics or driving requires application-specific testing, failure detection [Chen et al., 2025b], uncertainty estimation [Chen et al., 2026], and validated fallback mechanisms.

Acknowledgments and Disclosure of Funding

This research is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 3 Grant (MOE-MOET32022-0001). The authors declare no competing interests.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tim Brödermann, David Bruggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc Van Gool. Muses: The multi-sensor semantic perception dataset for driving under uncertainty. In *European Conference on Computer Vision*, pages 21–38. Springer, 2024.

Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4159–4166. IEEE, 2023.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7088–7097, 2021.

Jialei Chen, Xu Zheng, Danda Pani Paudel, Luc Van Gool, Hiroshi Murase, and Daisuke Deguchi. Bixformer: A robust framework for maximizing modality effectiveness in multi-modal semantic segmentation. *arXiv preprint arXiv:2506.03675*, 2025a.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Tingting Chen, Beibei Lin, Srinivas Anumasa, Vedant Shah, Zifeng Yuan, Qiran Zou, Anirudh Goyal, and Dianbo Liu. Auto-Discovery-Bench: Diagnosing structured state tracking in oracle-guided discovery. *arXiv preprint arXiv:2502.15224*, 2025b.

Tingting Chen, Beibei Lin, Zifeng Yuan, Qiran Zou, Hongyu He, Anirudh Goyal, Yew-Soon Ong, and Dianbo Liu. HypoSpace: A diagnostic benchmark for set-valued hypothesis generation under underdetermination and sublinear coverage bounds. In *International Conference on Machine Learning (ICML)*, 2026.

Shujian Deng, Xin Zhang, Wen Yan, Eric I-Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. Deep learning in digital pathology image analysis: A survey. *Frontiers of Medicine*, 14(4):470–487, 2020.

Wanjun Du, Zifeng Yuan, Tingting Chen, Fucai Ke, Beibei Lin, and Shunli Zhang. WeatherReasonSeg: A benchmark for weather-aware reasoning segmentation in visual language models. In *European Conference on Computer Vision (ECCV)*, 2026.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022.

Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13630–13639, 2023.

Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. *arXiv preprint arXiv:2406.01210*, 2024.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

Zexi Jia, Kaichao You, Weihua He, Yang Tian, Yongxiang Feng, Yaoyuan Wang, Xu Jia, Yihang Lou, Jingyi Zhang, Guoqi Li, et al. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32:1829–1842, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. *arXiv preprint arXiv:2408.01343*, 2024.

Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21694–21704, 2023.

Xiangtai Li, Tao Zhang, Yanwei Li, Haobo Yuan, Shihao Chen, Yikang Zhou, Jiahao Meng, Yueyi Sun, Shilin Xu, Lu Qi, Tianheng Cheng, Yi Lin, Zilong Huang, Wenhao Huang, Jiashi Feng, and Guang Shi. DenseWorld-1M: Towards detailed dense grounded caption in the real world. *arXiv preprint arXiv:2506.24102*, 2025a.

Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17182–17191, 2022.

Ziqi Li, Tao Gao, Yisheng An, Ting Chen, Jing Zhang, Yuanbo Wen, Mengkun Liu, and Qianxi Zhang. Brain-inspired spiking neural networks for energy-efficient object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3552–3562. IEEE, 2025b.

Ziqi Li, Tao Gao, Shutao Li, Ting Chen, Yisheng An, Yuanbo Wen, and Tao Lei. Frequency-oriented adaptive real-time object detector for cluttered traffic scenes. *Nature Communications*, 17(1):9787, 2026. doi: 10.1038/s41467-026-76346-1.

Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19800–19808, 2022.

Chenfei Liao, Xu Zheng, Yuanhuiyi Lyu, Haiwei Xue, Yihong Cao, Jiawen Wang, Kailun Yang, and Xuming Hu. Memorysam: Memorize modalities and semantics with segment anything model 2 for multi-modal semantic segmentation. *arXiv preprint arXiv:2503.06700*, 2025.

Beibei Lin, Zifeng Yuan, and Tingting Chen. RGB-to-polarization estimation: A new task and benchmark study. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Jiyuan Liu, Jia Lin, Xiaofei Zhou, Runmin Cong, Deyang Liu, and Zhi Liu. M4-SAM: Multi-modal mixture-of-experts with memory-augmented SAM for RGB-D video salient object detection. *arXiv preprint arXiv:2605.11760*, 2026.

Ruiping Liu, Jiaming Zhang, Kunyu Peng, Yufan Chen, Ke Cao, Junwei Zheng, M Saquib Sarfraz, Kailun Yang, and Rainer Stiefelhagen. Fourier prompt tuning for modality-incomplete scene segmentation. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 961–968. IEEE, 2024.

Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1020–1030, 2024.

Shishir Muralidhara, Didier Stricker, and René Schuster. MILE: Mixture of incremental LoRA experts for continual semantic segmentation across domains and modalities. *arXiv preprint arXiv:2605.03555*, 2026.

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.

Valentinos Pariza, Mohammadreza Salehi, Gertjan Burghouts, Francesco Locatello, and Yuki M Asano. Neco: Improving dinov2's spatial representations in 19 gpu hours with patch neighbor consistency. *arXiv e-prints*, pages arXiv–2408, 2024.

Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 4980–4989, 2017.

Xinyang Pu, Hecheng Jia, Linghao Zheng, Feng Wang, and Feng Xu. Classwise-sam-adapter: Parameter efficient fine-tuning adapts segment anything to sar domain for semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Ziniu Qian, Zihua Wang, Xin Zhang, Bingzheng Wei, Maode Lai, Jianzhong Shou, Yubo Fan, and Yan Xu. MSNSegNet: Attention-based multi-shape nuclei instance segmentation in histopathology images. *Medical & Biological Engineering & Computing*, 62(6):1821–1836, 2024.

Ukcheol Shin, Kyunghyun Lee, In So Kweon, and Jean Oh. Complementary random masking for RGB-Thermal semantic segmentation. In *IEEE International Conference on Robotics and Automation*, 2024.

Mengke Song, Wenfeng Song, Guowei Yang, and Chenglizhao Chen. Improving rgb-d salient object detection via modality-aware decoder. *IEEE Transactions on Image Processing*, 31:6124–6138, 2022.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Zhiqiang Teng, Tingting Chen, Beibei Lin, Zifeng Yuan, Xuanyi Li, Xuanyu Zhang, and Shunli Zhang. RaindropGS: A benchmark for 3D gaussian splatting under raindrop conditions. *arXiv preprint arXiv:2510.17719*, 2025.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 216–226. Springer, 2023.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024.

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33: 4835–4845, 2020.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195, 2022.

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*, pages 28619–28630, 2024.

Lekang Wen, Jing Xiao, Liang Liao, Jiajun Chen, and Mi Wang. CHARM: Collaborative harmonization across arbitrary modalities for modality-agnostic semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10603–10611, 2026. doi: 10.1609/aaai.v40i13.38033.

Wei Wu, Tao Chu, and Qiong Liu. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. *Pattern Recognition*, 131:108881, 2022.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of LoRA experts. *arXiv preprint arXiv:2404.13628*, 2024.

Ruihao Xia, Chaoqiang Zhao, Meng Zheng, Ziyan Wu, Qiyu Sun, and Yang Tang. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21572–21581, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Weilong Yan, Xin Zhang, and Robby T. Tan. ER-LoRA: Effective-rank guided adaptation for weather-generalized depth estimation. *arXiv preprint arXiv:2509.00665*, 2025.

Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision*, pages 677–695. Springer, 2022.

Xin Yang, Xin Zhang, and Xinchao Wang. ERF: A benchmark dataset for robust semantic segmentation under extreme rainfall conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9301–9309, 2025.

Zhiwen Yang, Jiayin Li, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-RWKV: Efficient and effective medical image restoration with RWKV. *IEEE Journal of Biomedical and Health Informatics*, 30(1):513–526, 2026a. doi: 10.1109/JBHI.2025.3588555.

Zhiwen Yang, Yang Zhou, Haowei Chen, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. UniPET: A universal network for high-quality PET image denoising across varied dose reduction factors. *Medical Image Analysis*, 111:104059, 2026b. doi: 10.1016/j.media.2026.104059.

Bo-Wen Yin, Jiao-Long Cao, Xuying Zhang, Yuming Chen, Ming-Ming Cheng, and Qibin Hou. Omnisegmentor: A flexible multi-modal learning framework for semantic segmentation. *arXiv preprint arXiv:2509.15096*, 2025.

Xiaowen Ying and Mooi Choo Chuah. Uctnet: Uncertainty-aware cross-modal transformer network for indoor rgb-d semantic segmentation. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with MLLM for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a. doi: 10.1109/TPAMI.2026.3720379.

Zifeng Yuan, Tingting Chen, Dewen Zhang, Yuan Gao, Wenkai Shan, Beibei Lin, and Aaron Danner. Tailored polarization-switchable VCSEL arrays for photonic Ising computing. *Applied Physics Letters*, 127(22):221102, 2025a.

Zifeng Yuan, Dewen Zhang, Yuan Gao, Luo Qi, Wujie Fu, and Aaron Danner. Large-scale fabrication and analysis of polarization behavior in VCSELs with tailored apertures. *Journal of Lightwave Technology*, 43(14):6819–6827, 2025b.

Zifeng Yuan, Dewen Zhang, Lei Shi, Yutong Liu, and Aaron Danner. Enhanced polarization locking in VCSELs. *Applied Physics Letters*, 126(15):151101, 2025c.

Zifeng Yuan, Beibei Lin, Yutong Liu, Wujie Fu, Yuan Gao, Tingting Chen, and Aaron Danner. Drive-efficient GHz-rate binary-state updates enabled by VCSEL polarization switching. *Laser & Photonics Reviews*, page e71890, 2026b.

Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. Issafe: Improving semantic segmentation in accidents by fusing event-based data. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1132–1139. IEEE, 2021a.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 2023a.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023b.

Jingwei Zhang, Xin Zhang, Ke Ma, Rajarsi Gupta, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. Gigapixel whole-slide images classification using locally supervised learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 192–201. Springer Nature Switzerland, 2022.

Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2021b.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. OMG-LLaVA: Bridging image-level, object-level, pixel-level reasoning and understanding. In *Advances in Neural Information Processing Systems*, volume 37, pages 71737–71767. Curran Associates, Inc., 2024a. doi: 10.52202/079017-2291.

Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5404–5413, 2023c.

Xin Zhang and Ying-Cong Chen. Adaptive domain generalization via online disagreement minimization. *IEEE Transactions on Image Processing*, 32:4247–4258, 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*, pages 14527–14537, 2025.

Xin Zhang, Weixuan Kou, Eric I-Chao Chang, He Gao, Yubo Fan, and Yan Xu. Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device. *Computers in Biology and Medicine*, 103:71–81, 2018.

Xin Zhang, Jinheng Xie, Yuan Yuan, Michael Bi Mi, and Robby T. Tan. HEAP: Unsupervised object discovery and localization with contrastive grouping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7323–7331, 2024b.

Xin Zhang, Haochen Wang, Yikang Zhou, Zhuochen Wang, Xiangtai Li, and Robby T. Tan. Actor as its own critic: Unifying region understanding and localization via CycleGRPO. In *European Conference on Computer Vision (ECCV)*, volume 17044 of *Lecture Notes in Computer Science*, pages 586–604. Springer, 2026. doi: 10.1007/978-3-032-37577-3_32.

Xu Zheng, Yuanhuiyi Lyu, Lutao Jiang, Jiazhou Zhou, Lin Wang, and Xuming Hu. Magic++: Efficient and resilient modality-agnostic semantic segmentation via hierarchical modality selection. *arXiv preprint arXiv:2412.16876*, 2024a.

Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning modality-agnostic representation for semantic segmentation from any modalities. In *European Conference on Computer Vision*, pages 146–165. Springer, 2024b.

Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation. In *European Conference on Computer Vision*, pages 192–212. Springer, 2024c.

Xu Zheng, Yuanhuiyi Lyu, Lutao Jiang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Reducing unimodal bias in multi-modal semantic segmentation with multi-scale functional entropy regularization. *arXiv preprint arXiv:2505.06635*, 2025.

Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, Shunping Ji, Anran Wang, Zhuochen Wang, Yujing Wang, Cheng Chen, and Xiangtai Li. SAMTok: Representing any mask with two words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 37852–37863, 2026.

Chenyang Zhu, Bin Xiao, Lin Shi, Shoukun Xu, and Xu Zheng. Customize segment anything model for multi-modal semantic segmentation with mixture of lora experts. *arXiv preprint arXiv:2412.04220*, 2024.

Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021.

A Controlled Comparisons and Experimental Protocols

We separate external same-backbone references from controls trained within our framework. For MUSES MiT-B0/B2, internal controls use a frozen encoder, SegFormer decoder, the same split, augmentations, optimizer, 200-epoch schedule, 1024×1024 inputs, rank-4 adaptation where applicable, EMA-query PMM, effective global batch size 16, and the best full-modality validation checkpoint selected independently for each run. Seeds are 3407, 3408, and 3409 for repetitions and 3407 for individual controls. These experiments test backbone generality on the evaluated hierarchical MiT scales, not arbitrary architectures.

Deep Prompt inserts 64 modality-specific learnable prompts into every Transformer block. Shared LoRA applies one rank-4 Q/V and FFN update shared across sensors. Adapter inserts modality-specific width-32 residual bottlenecks in stages 2–4. Separate LoRA uses private Q/V and FFN residual branches without cross-modal routing. Full FuseAdapt adds receiver-conditioned FFN residual exchange. All variants average the resulting visible features before the shared decoder.

Table 9: MUSES full-modality mIoU with MiT-B0. External references share the backbone family; internal controls additionally share the decoder and recipe. Trainable parameter counts include optimized components of the internal setup.

Method	Comparison	Trainable (M)	mIoU
MAGIC	External	3.72	49.02
Any2Seg	External	3.72	50.00
Deep Prompt	Internal	2.95	49.47
Shared LoRA	Internal	2.81	53.56
Adapter	Internal	2.95	54.80
FuseAdapt	Internal	2.94	**57.54**

Table 10: MUSES full-modality mIoU using MiT-B2. Any2Seg has no reported MUSES/MiT-B2 result and is omitted.

External method	mIoU	Internal variant (seed 3407)	mIoU
CMNeXt	58.28	Shared	65.38
MAGIC	55.36	Separate	66.96
StitchFusion	53.40	FuseAdapt	**70.60**
CHARM [Wen et al., 2026]	59.36		

DINOv2 capacity. Under the same MUSES split, seed 3407, 200 epochs, rank 4, EMA-query PMM, optimizer schedule, and effective batch size 16, DINOv2-S/B achieve 77.28/80.16 full-modality best-validation mIoU. The 2.88-point difference confirms the contribution of backbone capacity. DINOv2-S was selected to keep the original 3.29M trainable budget comparable to MAGIC/Any2Seg (3.72M each); it does not match their backbone family.

Fusion controls on MCubeS. All rows in Table 11 use DINOv2-B and a shared segmentation decoder. Frozen mean averages frozen per-modality features without LoRA. Frozen concatenation concatenates those features and projects them through a trainable 1×1 layer at each scale. Token-attention fusion uses eight-head attention across sensor tokens at each common-grid location. Late-logit fusion uses private rank-8 Q/V and FFN updates without routing, decodes each stream independently with the shared head, and averages logits. Shared and Separate use shared or private low-rank branches, respectively, followed by feature averaging.

Information-flow alternatives. All-path routed LoRA routes the low-rank Q/V residuals as well as the FFN residuals, testing whether attention adaptation should also mix across sensors. Dense-feature routing preserves private Q/V but exchanges complete FFN features, testing whether the low-rank residual restriction matters. The MUSES MiT-B0 controls share rank, PMM, training recipe, effective batch size, seed, trainable parameter count, and checkpoint rule. Table 7 also reports the corresponding MCubeS/DINOv2-B results. These controlled alternatives isolate architectural choices rather than reproducing StitchFusion or MLE-SAM in full.

Table 11: MCubeS, all four modalities (RADN): fusion/adaptation mIoU.

Method	RADN mIoU
Frozen mean	51.16
Frozen concatenation	52.42
Token-attention fusion	52.79
Late-logit fusion	54.55
Shared	54.69
Separate	56.09
FuseAdapt	**57.62**

B Multi-Seed Reliability

Table 12 reports all runs in the MUSES MiT comparison. Each run is trained for the same budget and evaluated using its validation-selected full-modality checkpoint. For values x_1, x_2, x_3, we report $\bar{x} = \frac{1}{3}\sum_i x_i$ and sample standard deviation $s = \sqrt{\sum_i (x_i - \bar{x})^2/(3-1)}$. This measures run-to-run variation under a fixed split and recipe; it is not a standard error or confidence interval. We do not infer formal significance from three seeds. Full exceeds Shared in all six paired comparisons: by

Table 12: MUSES full-modality mIoU for each random seed and mean $\pm$ sample SD.

Backbone	Variant	3407	3408	3409	Mean $\pm$ SD
MiT-B0	Shared	53.56	54.74	54.09	54.13 ± 0.59
	Separate	54.31	56.34	56.94	55.86 ± 1.38
	Full	57.54	57.77	57.48	57.60 ± 0.15
MiT-B2	Shared	65.38	65.69	66.32	65.80 ± 0.48
	Separate	66.96	68.29	66.95	67.40 ± 0.77
	Full	70.60	70.67	70.58	70.62 ± 0.05

3.98/3.03/3.39 on MiT-B0 and 5.22/4.98/4.26 on MiT-B2. The mean improvements over Shared are 3.47/4.82, and over Separate 1.74/3.22. In a separate three-seed fixed-checkpoint evaluation, the parameter-matched Shared control and Full obtain 54.52 ± 0.29 and 57.41 ± 0.27 on MCubeS, and 76.63 ± 0.43 and 77.17 ± 0.38 on MUSES with DINOv2-S. The latter gain of 0.54 is modest relative to variation; the clearer MiT margins should not be generalized to all backbones. A stronger Shared representation may leave less room for routing, but this is an interpretation rather than a causal result.

C PMM Objective Controls and Teacher Reliability

The MCubeS MISS comparisons keep the backbone, decoder, modality masking, segmentation supervision, and training schedule fixed. Output KD matches partial-input student logits to full-context EMA teacher logits through pixelwise KL divergence. Global-feature KD matches the student's fused representation to the full-context EMA fused representation using cosine loss. Neither identifies a missing target sensor. Per-token regression instead predicts its latent tokens at the same spatial indices. Query without EMA uses the same query predictor as PMM but a detached online full-context target. Full PMM uses an EMA target and target-specific query cross-attention. An EMA teacher without a matching loss would not provide student supervision.

PMM exceeds output KD by 3.05 mean mIoU and fused-feature KD by 2.32. It improves the former in all 15 subsets and the latter in 14/15. Linear prediction improves dropout-only training by 6.04; query prediction improves the linear control by 3.18; EMA improves the no-EMA query by 2.30. These results support modality-conditioned latent prediction beyond a generic consistency objective, without claiming masking or EMA as new mechanisms.

C.1 Full-context targets are an average prior

Teacher and student share the same frozen initialization and architecture. Only student adaptation parameters receive gradients, and teacher adaptation parameters follow the EMA update with coefficient 0.999. PMM uses detached teacher representations rather than segmentation pseudo-labels.

Table 13: MCubeS MISS mIoU for all objective controls. Mean averages all 15 nonempty subsets, including RADN. R/A/D/N denote RGB/AoLP/DoLP/NIR.

Subset	Dropout	Output KD	Linear	Query, no EMA	Feature KD	PMM
R	51.02	51.40	51.38	53.97	53.46	**54.96**
A	13.60	34.92	31.08	37.99	**38.66**	37.60
D	22.47	46.13	42.02	44.75	43.70	**47.14**
N	33.61	45.79	44.97	46.67	45.65	**49.12**
RA	53.16	52.08	53.64	55.10	54.27	**56.17**
RD	54.92	53.99	54.64	54.57	55.06	**57.23**
RN	54.04	52.62	54.82	53.06	53.64	**56.29**
AD	26.37	45.39	43.36	46.35	44.64	**47.96**
AN	38.26	48.33	47.38	48.88	48.66	**51.65**
DN	40.24	49.63	49.62	49.67	49.61	**52.74**
RAD	55.33	53.70	55.01	54.86	55.40	**57.19**
RAN	55.67	53.36	55.31	53.63	54.09	**56.80**
RDN	56.61	54.71	56.06	54.52	55.47	**57.58**
ADN	43.02	50.56	50.15	50.32	50.75	**53.30**
RADN	56.66	55.00	56.20	54.58	55.55	**57.62**
Mean	43.67	49.84	49.71	50.59	50.57	**52.89**

Full-modality input does not guarantee better predictions when a sensor is degraded. Over 1,428 MCubeS image/subset pairs (102 images and 14 proper subsets), full-input predictions improve per-image mIoU by 4.44 on average and are better in 81.72% of pairs. Thus full context is useful on average under this distribution, but it is not a per-sample oracle. Reliability-aware target weighting under sensor corruption remains future work.

D Degradation and Spatial-Shift Evaluation

Degraded but present modalities. Table 14 evaluates all four DELIVER modalities (R-D-E-L) on the benchmark's predefined weather and sensor-degradation subsets. MB/OE/UE/LJ/EL denote motion blur, over-exposure, under-exposure, LiDAR jitter, and event low resolution. This is not an additional arbitrary-corruption benchmark. FuseAdapt leads eight conditions and the mean, while Any2Seg leads Night and UE.

Table 14: DELIVER, four present modalities: mIoU by predefined condition.

Condition	HRFuser	CMNeXt	Any2Seg	FuseAdapt
Cloudy	56.20	68.70	69.19	**69.22**
Foggy	52.39	65.67	67.64	**67.93**
Night	49.85	62.46	**67.25**	66.44
Rainy	52.53	67.50	67.00	**68.76**
Sunny	54.02	66.57	68.25	**68.93**
MB	49.44	62.91	66.23	**66.39**
OE	46.31	64.59	64.89	**67.21**
UE	46.92	60.00	**65.69**	61.39
LJ	53.94	65.92	69.05	**69.08**
EL	52.72	65.48	66.73	**68.23**
Mean	51.43	64.98	67.19	**67.36**

Latent prediction under spatial shifts. We mask each MCubeS target sensor in turn, then jointly shift the other three visible sensors by 0/2/4/8/16 input pixels relative to the clean full-context EMA target. Visible–visible alignment is preserved while target–visible correspondence changes. We compute channel-wise cosine between predicted target tokens and their teacher counterparts, averaged over tokens, 102 images, four target modalities, and three seeds. At 0 pixels the target is still missing and visible inputs are unshifted. Query prediction has the smaller drop in all 12 seed–target comparisons at every nonzero shift; its 16-pixel drop is approximately 32 times smaller.

Table 15: MCubeS predicted/teacher latent cosine under target–visible shifts. Final column reports mean drop ± SD as reported across the three seeds.

Predictor	0 px	2 px	4 px	8 px	16 px	Drop, 0–16 px
Linear	.86968	.86766	.86583	.86136	.85332	$.01636 \pm .00323$
Query	.74025	.74012	.74008	.74006	.73974	$.00051 \pm .00029$

Linear prediction has higher absolute clean cosine. The experiment measures sensitivity of latent prediction to these translations, not general reconstruction superiority or end-to-end robustness to arbitrary native token lengths, fields of view, or temporal offsets.

E End-to-End Inference Efficiency

We profile all methods on the same idle H100 at 1024×1024, batch size 1, FP32, with 20 warm-up iterations and 100 measured runs. FuseAdapt, CMNeXt, and StitchFusion use MiT-B0 and 256-dimensional segmentation decoders. Any2Seg and MAGIC use MiT-B0 with their paper-specific heads. Deployed parameters count unique parameters retained for inference, counting shared backbones once and removing PMM's EMA teacher and query predictor. An asterisk identifies an inference graph reconstructed from its paper because executable code was unavailable at the time of profiling; the dagger denotes the released DFormer-L+LightHam proxy for OmniSegmentor.

Table 16: End-to-end inference cost under the common H100 protocol. Rows list one, two, or three input modalities. Latency is ms, throughput is images/s, memory is peak GiB, and parameters are deployed millions.

Method	Modalities	Latency	Throughput	Memory	Params
CMNeXt	1	7.30	136.96	0.70	3.72
	2	18.45	54.19	0.74	10.29
	3	19.63	50.93	0.75	10.30
StitchFusion	1	8.32	120.26	0.70	4.26
	2	16.41	60.93	0.72	4.39
	3	26.69	37.47	0.73	4.56
Any2Seg*	1	6.53	153.08	0.60	3.72
	2	10.26	97.43	1.16	3.72
	3	14.69	68.09	1.71	3.72
MAGIC*	1	6.49	154.16	0.60	3.72
	2	10.25	97.53	1.16	3.79
	3	14.72	67.92	1.71	3.85
FuseAdapt	1	9.40	106.33	0.73	3.78
	2	15.87	63.02	1.18	3.84
	3	21.05	47.50	1.75	3.91
MLE-SAM*	1	33.91	29.49	0.95	89.04
	2	61.95	16.14	1.06	89.04
	3	89.72	11.15	1.17	89.04
MemorySAM	1	35.56	28.12	0.87	84.08
	2	78.39	12.76	1.21	84.08
	3	125.23	7.99	1.43	84.08
OmniSegmentor†	2	45.48	21.99	0.93	39.02

FuseAdapt's three-input latency is close to CMNeXt and below StitchFusion in this profile; its unique deployed parameter count is smaller than both. However, Any2Seg/MAGIC are faster, and CMNeXt/StitchFusion use less peak memory. Reconstructed graphs and the OmniSegmentor proxy should not be interpreted as official implementation benchmarks. We omit traced FLOPs because unsupported routing, normalization, and interpolation operations cause method-dependent

undercounting. Small adaptation-module overhead should not be confused with the total cost of each additional frozen-encoder stream.

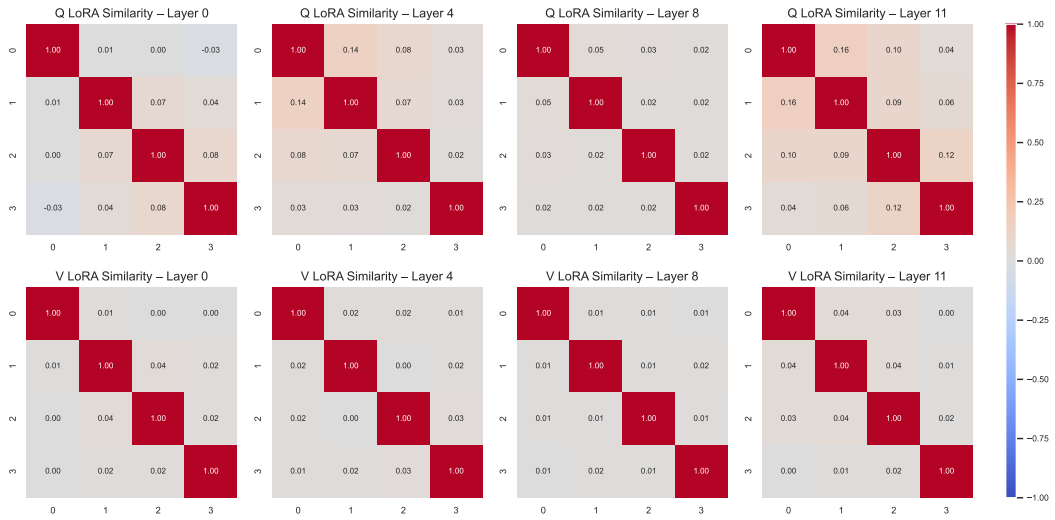


Figure 5: Cosine similarity matrices of modality-specific LoRA updates for the Q-projection (top row) and V-projection (bottom row) across different transformer layers. Each heatmap shows the pairwise cosine similarity between the LoRA updates ΔW of the four modalities of DELIVER. Across all layers, different modalities exhibit consistently low pairwise cosine similarity, indicating that the LoRA updates for each modality occupy distinct and nearly orthogonal subspaces. The indices 0, 1, 2, and 3 correspond to the four input modalities.

F Methodological Justification

F.1 Rationale for Using Modality-Specific LoRA in Self-Attention Adaptation.

To examine whether self-attention should apply a fused LoRA module across modalities or maintain modality-specific adaptation paths, we analyze the learned LoRA updates ΔW across both the q and v projections for each modality on two datasets, DELIVER and MCubeS. For each projection, we also evaluate shallow, mid, and deep transformer layers and report cosine-similarity matrices.

As shown in Fig. 5 and Fig. 6, across all layers and for both projections, the LoRA updates of different modalities exhibit consistently low pairwise cosine similarity, forming near-orthogonal subspaces rather than converging toward a shared direction. This indicates that each modality learns distinct and highly specialized correction directions within the attention mechanism.

Low pairwise similarity supports distinct sensor-specific correction directions, but does not by itself establish that routing them is harmful. The all-path routing comparison in Table 7 tests that choice directly. Keeping Q/V private retains each sensor's attention rule while the FFN residual path permits cross-modal communication.

F.2 Comparison with Mixture-of-Experts (MoE)

Each FuseAdapt low-rank branch has a fixed source-sensor identity, while receiver-conditioned routing weights vary by token and input. All available sensor branches participate without top-k selection or a load-balancing loss; absent sensors are masked. This distinguishes branch assignment from dynamic communication and does not imply fixed mixing weights. A sparse MoE-style LoRA control obtains 54.94 versus 57.62 mIoU on MCubeS. The all-path and dense-feature controls in Table 7 more directly test which activations should cross sensor streams.

F.3 Details of PMM.

After the full-input warm-up, with probability 0.5 we retain all modalities; otherwise we draw independent Bernoulli-0.5 visibility masks and enforce a nonempty visible set per sample. A target is sampled only when at least one modality is missing. The teacher always receives the full input;

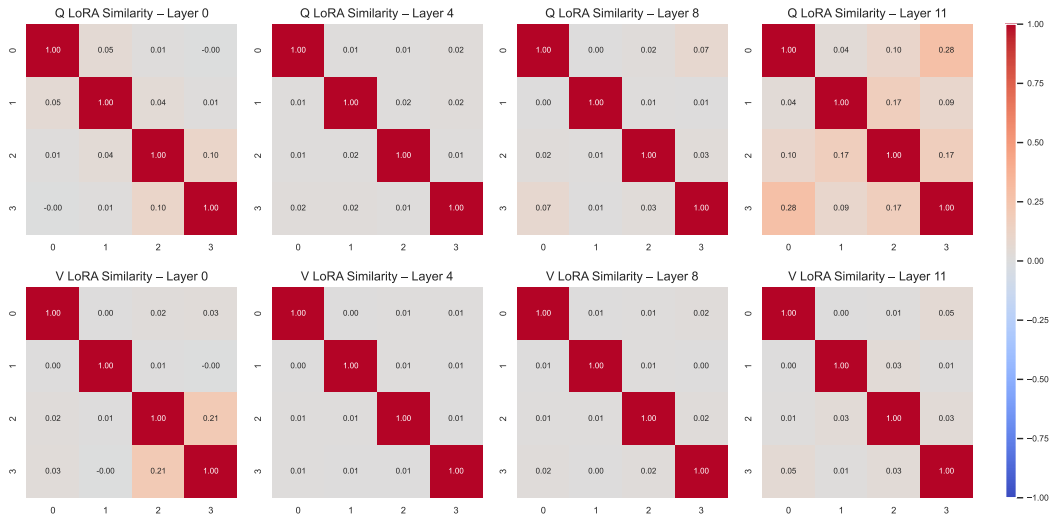


Figure 6: Cosine similarity matrices of modality-specific LoRA updates for the Q-projection (top row) and V-projection (bottom row) across different transformer layers. Each heatmap shows the pairwise cosine similarity between the LoRA updates ΔW of the four modalities in MCubeS. As with DELIVER, the cross-modal LoRA directions remain consistently low in similarity across all layers. The indices 0, 1, 2, and 3 correspond to the four input modalities.

its designated target-modality representation is detached. Modality-specific query sets attend to visible features and produce the same number of tokens as the target. For full-input samples, the missing-target PMM term is zero. Inference uses the actually available subset: all modalities for MCSS and the evaluated subset for MISS. The teacher and predictor are discarded.

F.4 Class-balanced OHEM Cross-Entropy Loss.

To better handle class imbalance and focus optimization on informative pixels, we adopt a weighted Online Hard Example Mining (OHEM) strategy with per-batch adaptive class weights, based on the vanilla OHEM implemented in DELIVER. Given the prediction $\mathbf{P} \in \mathbb{R}^{B\times C\times H\times W}$ and ground-truth labels $\mathbf{Y} \in \mathbb{R}^{B\times H\times W}$, the pixel-wise cross-entropy loss is first computed as:

$$\mathcal{L}_{\text{CE}}(i) = -\sum_{c=1}^{C} w_c \cdot \mathbf{1}[y_i = c] \cdot \log P_{i,c}, \tag{13}$$

where w_c denotes the class-specific weight for class c, computed per batch using the inverse frequency of observed labels, clipped to a maximum value of w_{max}:

$$w_c = \begin{cases} \min\left(\dfrac{1}{f_c+\epsilon} \Big/ \min\limits_{c'} \dfrac{1}{f_{c'}+\epsilon},\ w_{\text{max}}\right), & \text{if } f_c > 0 \\ w_{\text{default}}, & \text{otherwise} \end{cases} \tag{14}$$

where f_c is the pixel count of class c in the current batch and w_{default} is the fallback weight for absent classes. w_{max} is set to 10, and w_{default} is set to 1.

To prioritize informative samples, we follow OHEM and select pixels with loss greater than a threshold τ (converted to logit space as $-\log\tau$). Let $\mathcal{H}$ be the set of such hard pixels (τ is set to 0.7):

$$\mathcal{H} = \{i \mid \mathcal{L}_{\text{CE}}(i) > -\log\tau\} \tag{15}$$

If $|\mathcal{H}| < N_{\text{min}}$, we select the top N_{min} hardest pixels by loss. The final loss is the mean over these hard pixels:

$$\mathcal{L}_{\text{OHEM}} = \frac{1}{|\mathcal{H}'|} \sum_{i \in \mathcal{H}'} \mathcal{L}_{\text{CE}}(i), \tag{16}$$

where $\mathcal{H}' = \mathcal{H}$ if $|\mathcal{H}| \geq N_{\text{min}}$, otherwise it is set to the top N_{min} pixels. Both the loss weights of the segmentation and PMM are set to 1.

Algorithm 1 PMM training after full-input warm-up

1: Sample nonempty visible set $\mathcal{V}$ using the full/masked mixture above.
2: Compute student features with routing normalized over $\mathcal{V}$ and segmentation loss $\mathcal{L}_{\text{seg}}$.
3: Set $\mathcal{L}_{\text{pmm}} = 0$.
4: **if** $\mathcal{V} \neq \mathcal{M}$ **then**
5: Sample $t \in \mathcal{M} \setminus \mathcal{V}$ uniformly.
6: Compute detached target y_t with the full-context EMA teacher.
7: Predict $\hat{z}_t = g_\phi(Q_t, \hat{\mathbf{X}}_{\mathcal{V}})$ and compute cosine loss $\mathcal{L}_{\text{pmm}}$.
8: **end if**
9: Update student and predictor using $\mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{pmm}}$.
10: Update teacher: $\theta_\tau \leftarrow 0.999\theta_\tau + 0.001\theta_s$.

G Implementation Details

We train the proposed method on 8 A5000 GPUs for DELIVER, 4 for MUSES, and 4 for MCubeS. Initial learning rates are 3.5e^{-4} (MCubeS), 4e^{-4} (MUSES) and 4e^{-4} (DELIVER), with a 10-epoch warm-up at 0.1×. Training is conducted for 500, 200, and 200 epochs on MCubeS, DELIVER, and MUSES respectively. Each run begins with a warm-up stage (200 epochs for MCubeS and 15 epochs for both DELIVER and MUSES), during which the model is trained using all modalities without any masking. We use AdamW with batch sizes of 1 (DELIVER) and 4 (MUSES, MCubeS) per GPU. Input resolutions are 1024×1024 (DELIVER, MUSES) and 512×512 (MCubeS), with standard augmentations: random resizing (0.5–2.0), flipping, color jitter, blur, and cropping. FuseAdapt uses rank 8 for DELIVER and MCubeS, and 4 for MUSES. DINOv2-B (for DELIVER and MCubeS) [Oquab et al., 2023] and DINOv2-S (MUSES) serve as the frozen VFMs. Standard LoRA with the low-rank matrices A and B such that A is initialized with a small random Gaussian (e.g., $\mathcal{N}(0, 0.01)$) and B is initialized as zero, ensuring the initial weight perturbation is zero. The number of learnable tokens in the predictor is the same as the sequence length of the target feature.

H More Ablation Studies

Rank r We evaluate the performance of the proposed method with different rank r. The experiments are conducted on MCubeS on the MCSS setting. As shown in Tab. 17, FuseAdapt significantly improves performance over a frozen VFM, with the best results obtained at rank 8.

Table 17: Results of different rank r under MCSS setting on MCubeS. All models are trained with four modalities.

r	0	4	8	16
RADN	51.16	55.86	**57.62**	57.03

Number of Layers Equipped with FuseAdapt LoRAs We further evaluate the impact of the insertion frequency of FuseAdapt LoRAs in the VFM, where a module is inserted every N layers. A larger N indicates fewer inserted LoRAs. The experiments are conducted on MCubeS under the MCSS setting. As shown in Tab. 18, inserting more FuseAdapt LoRAs modules yields greater performance improvements.

Routing Fusion in FFNs We introduce independent, modality-specific routing modules at the FFN stage to enable controlled and dynamic cross-modal fusion. Unlike attention layers that focus

Table 18: Results of inserting FuseAdapt LoRAs every N layers under MCSS setting on MCubeS. All models are trained with four modalities.

N	0	1	2	3
RADN	51.16	**57.62**	56.38	55.87

on intra-modal context, these routing blocks refine each modality's representation by selectively integrating complementary signals from others, based on the current context. This fusion occurs after independent encoding, preventing early entanglement and allowing the model to adaptively weight modalities depending on availability, especially beneficial under modality masking. Importantly, each modality has its own dedicated routing layer, which maintains structural disentanglement and minimizes interference from masked or undertrained modalities.

The routing module adopts a lightweight noisy gating mechanism, adapted from sparsely activated Mixture-of-Experts (MoE) designs. Given hidden states, the router outputs modality-specific routing scores through a linear layer, and injects learned Gaussian noise via a second linear layer followed by softplus. The input size matches the feature dimension of the VFM output, while the output size corresponds to the number of training modalities. The resulting noisy logits enable soft, differentiable expert selection. Routers are trained end-to-end as part of the main optimization, without any additional loss terms.

We evaluate the effectiveness of routing mechanisms. As shown in Table 19, incorporating routing fusion into the FFNs consistently improves performance. When no fusion is applied ("without routing"), the model processes each modality independently and therefore cannot benefit from cross-modal interactions. Using a simple average fusion yields only marginal gains, as it assumes that all modalities contribute equally and ignores their heterogeneous characteristics. In contrast, routing fusion adaptively selects and combines modality-dependent FFN experts, leading to a clear improvement (from 56.09 to 57.62 mIoU). This demonstrates that learned routing effectively captures the distinct roles that different modalities play in multi-modal feature refinement. We also observe that a shared router yields a competitive gain (57.24 mIoU), showing that routing is helpful even without modality-specific specialization.

Notably, unlike Mixture of Experts structures that perform token-level dynamic expert selection (e.g., top-K gating), our routing deterministically assigns each LoRA expert to a specific modality. All available modality branches participate without top-k selection or a load-balancing objective. Mixing weights remain token-dependent, and absent sources are masked before normalization.

Table 19: Results of various kinds of routing strategies under MCSS setting on MCubeS. All models are trained with four modalities.

	wo. routing	average	shared router	w. routing
RADN	56.09	56.43	57.24	**57.62**

Additional VFMs To assess backbone sensitivity, we evaluate our method with two alternative VFMs, EVA-CLIP-B [Sun et al., 2023] and NeCo-B [Pariza et al., 2024], on MCubeS (in MCSS setting with full-modality training). As shown in Table 20, the proposed approach maintains strong performance across all backbones:

Table 20: Results of applying various VFMs under MCSS setting on MCubeS. All models are trained with four modalities.

	EVA-CLIP	NeCo	DINOv2
RADN	56.02	56.28	**57.62**

Scalability of FuseAdapt We analyze the scalability of FuseAdapt under different backbones and input resolutions.

For DINOv2-B with the input size of 512×512 and rank of 8, adding an additional modality increases the parameter count by 1.11M, which corresponds to 1.12% of the full model. The memory cost,

including optimizer states under AdamW (in float32 type and without AMP), is 16.90 MB, and the computational cost is 4.8 GFLOPs, representing only 0.55% of the total model FLOPs (873.96 GFLOPs).

For DINOv2-S with the input size of 1024×1024 and rank of 4, each additional modality introduces 0.29M parameters (1.14% of the full model), 4.50 MB of memory, and 18.12 GFLOPs, which amount to 2.01% of the overall FLOPs (902.76 GFLOPs).

Although the cost of FuseAdapt grows linearly with the number of modalities, the relative overhead on parameters, memory, and computation remains very small. These figures describe adaptation-module overhead, not the total cost of processing an additional sensor through the frozen encoder. End-to-end deployment measurements are reported in Appendix E.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction have accurately reflected the paper's contribution and scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations in the Discussion and Conclusion.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not make theoretical contributions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The method, masking and EMA procedure, experimental settings, control definitions, and checkpoint-selection protocol are described in the main paper and appendices.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The benchmarks are publicly available. Code and reproduction instructions are provided through the project page (`https://devinxzhang.github.io/FuseAdapt-Page/`).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See the main paper and Appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix B reports per-seed results, means, and sample standard deviations across three independent seeds for the MUSES MiT-B0/B2 Shared/Separate/FuseAdapt comparison. The standard deviation uses denominator $n - 1$ and measures run-to-run variability, not a confidence interval. The other segmentation comparisons are reported from single runs, and no formal statistical-significance claim is made.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See the implementation details section.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses standard public multimodal semantic segmentation benchmarks. We do not collect new human-subject data, introduce personally identifiable information, or deploy the system in real-world decision-making settings. We follow the datasets' intended research use and discuss confidence reliability as a factor relevant to safer downstream perception.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The limitations discuss potential perception failures in robotics and driving, correlated sensor failures, and the need for uncertainty estimation, failure detection, and validated fallback mechanisms.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: To the best of our knowledge, the paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the used code, data, and models are properly cited and publicly available.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: All experiments rely exclusively on publicly released datasets. Consequently, no participant instructions, screenshots, or compensation details are applicable to this paper.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: All experiments use publicly released, de-identified datasets that were gathered under the ethics and IRB approvals of their original publishers. No additional participants were recruited and no personally identifiable information was processed. There were no participant risks to disclose and no new IRB review was required for this study.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core design of the method does not involve the LLMs.