ARCHITECTURAL MISSING-MODALITY ROBUSTNESS IN SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal semantic segmentation is typically evaluated with every modality present, but real-world deployment conditions often differ. A model trained on several modalities should therefore stay accurate on any non-empty subset of them at test-time. We propose **A**dapted **M**ixture **o**f **M**odalities (A-MoM), which—instead of relying on auxiliary signals, training strategies, or model-specific losses—attains robustness architecturally through: (1) token-level, context-dependent mixing, since a sensor's advantage is usually local; (2) an initialization that keeps mixing low by default, so the model does not learn to depend on all modalities; and (3) low-rank adapters that translate features between modality pairs. Across 4 datasets spanning 8 modalities, A-MoM improves the average robustness score by 1.5 and retains on average 92.7% of the mIoU of instances of A-MoM specialized to each modality-subset on MCubeS, while also improving the average all-modalities score by 0.7. Thus, A-MoM narrows the gap from research to deployment under real-world conditions.[1]

1 INTRODUCTION

A multimodal segmentation model should not fall apart when one of its sensors fails, and we show that 3 small changes to how a Transformer fuses modalities can increase its robustness.

Robots and autonomous vehicles must understand the scene around them before they can act safely in it. Semantic segmentation supplies much of that understanding, and it is well established that combining several sensors (e.g. RGB, event cameras, and depth) improves segmentation (Zhang et al., 2021), making multimodal segmentation an attractive choice for robot perception.

That improvement rests on an implicit assumption that every sensor the model is trained on is available every time the model is run. In practice, this cannot be guaranteed as sensors fail at the hardware and software levels, their measurements arrive asynchronously, and sensors degrade in adverse weather (e.g. fog and snow) and lighting conditions. Beyond outright failure, one would like to train a model once and deploy it across a fleet of robots and vehicles whose sensor suites differ. This is a step towards *train once, run anywhere* for resource-constrained models. Both cases require that a model trained on a set of modalities $\mathbb{M}$ remain accurate on any non-empty subset of $\mathbb{M}$ at test-time.

This kind of robustness is largely absent from the way fusion methods are evaluated. Performance is measured with all modalities present, and architectures are tuned accordingly, which rewards methods that tightly entangle the modalities. The consequences can be severe. For example, on MCubeS (Liang et al., 2022), when only near-infrared is available, CMNeXt (Zhang et al., 2023) merely recovers 3.0% of the mIoU of a model trained specifically on near-infrared, and GeminiFusion (Jia et al., 2024) falls to 4.3% in the same case. A score reported under full modality availability therefore says very little about what a model will do when sensors fail. We consequently report the robustness score as the average mIoU over all non-empty subsets of the training modalities alongside the usual all-modalities score, and we set out to discover a method that achieves high robustness.

We derive our method from 3 insights. First, a sensor's advantage is often local (Figure 1), meaning fusion should be at the token level and conditioned on context. Second, if modalities are fused at every opportunity, then the model learns to depend on all of them. Mixing should therefore be low by

[1]The code and model weights will be made publicly available upon acceptance.

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.

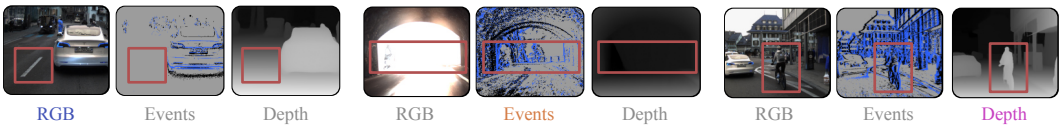

Figure 1: **Token-level motivation.** Sensor advantages are local, so the fusion decision belongs at the token level, illustrated by 1 case each for RGB, events, and depth.

default and only blend heavily where the model learns that it is beneficial. Third, modalities provide fundamentally different views, so the model should adapt features between each pair of modalities.

These insights lead us to propose A-MoM — *Adapted Mixture of Modalities*. We summarize our claims and contributions as follows:

❶ **Adapted Mixture of Modalities (A-MoM).** We identify 3 key insights and translate each into a design decision. Together, these design decisions result in A-MoM.

❷ **Robust to missing modalities.** We evaluate A-MoM on 4 datasets. A-MoM ranks highest in robustness score on all 4 datasets and improves the average robustness score by 1.5 over the next-best method, recovering on average 92.7% of the mIoU of instances of A-MoM specialized to each modality-subset on MCubeS. Additionally, A-MoM proves robust to adverse conditions with a 7.6 mIoU increase in foggy weather on MUSES (Brödermann et al., 2025a).

❸ **Accurate multimodal segmentation.** A-MoM remains strong when all modalities are present, improving the average overall score by 0.7 and obtaining the highest score on 2 of 4 datasets, or 3 of 4 datasets when GeminiFusion is reduced to a more comparable parameter budget.

2 Related Work

Missing-modality robustness. Prior work has shown that Transformers are sensitive to missing modalities (Ma et al., 2022). The literature proposes diverse methods to tackle this problem. One line of work conditions on external signals like language (Zheng et al., 2025a), weather (Brödermann et al., 2025b), and depth (Brödermann et al., 2026), on the premise that the external signal is useful for learning modality selection. Other works investigate training strategies (Lee et al., 2023; Liu et al., 2024; Cai et al., 2025), semi-supervised learning (Maheshwari et al., 2024), and parameter-efficient adaptation to all modality combinations (Reza et al., 2025). These methods are orthogonal to ours as we instead seek to understand missing-modality robustness from an *architectural* view. In this setting, prior works add necks and losses to a pretrained backbone (Li et al., 2025b; Zheng et al., 2025b; Wen et al., 2026). Our work differs by applying minimal changes to the pretrained backbone itself without any additional necks or losses.

Exchange-based models fuse modalities through pixel- or token-level exchanges of information between modalities. TokenFusion (Wang et al., 2022) proposes to replace features from pairs of modalities based on a learned score, while GeminiFusion (Jia et al., 2024) argues for soft replacement based on cross-attention. StitchFusion (Li et al., 2025a) exchanges information as a sum of pairwise adaptations of all modalities. Our work differs by studying missing-modality robustness explicitly, but it shares technical similarities with them, as A-MoM is also exchange-based: it generalizes TokenFusion using soft weighting of all modalities—not only pairs—but unlike GeminiFusion this decision does not involve computationally costly cross-attention. A-MoM also shares the idea of adapting features between modalities that StitchFusion pioneers.

Mixture of Experts (MoE) models learn to apply different expert subnetworks to each token (Fedus et al., 2022), and Han et al. (2024) investigate this method for fusion. A-MoM shares similarities with MoE models in name and in how it mixes multiple modalities and applies subnetworks between pairs of modalities. However, it differs in its motivation (missing-modality robustness and fusion) and technically by defining subnetworks as pairwise adaptations rather than specialized experts.

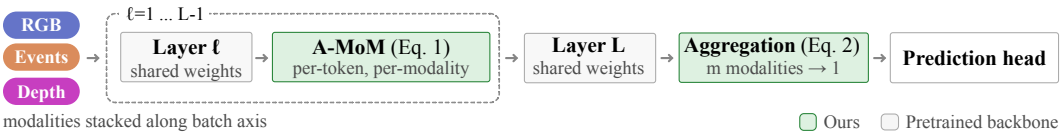

Figure 2: **A-MoM integration with a Transformer.** A single Transformer with shared weights processes all modalities stacked along the batch dimension; A-MoM fuses them between layers, independently at every token. A learned aggregation merges m tokens into 1 token at the stage output.

3 A-MoM: Adapted Mixture of Modalities

Problem formulation. We are given a set of m modalities $\mathbb{M}$. A sample, $\mathbb{A} = \{a_i \in \mathbb{R}^{h \times w \times c_i} : i \in \mathbb{M}\}$, consists of m images from different sensors that share a common height h and width w but may differ in their number of channels c_i. Our goal is a model that achieves high accuracy not only on the full sample $\mathbb{A}$, but also on any non-empty subset of modalities $\mathbb{M}_{miss} \subseteq \mathbb{M}$. In other words, the model should stay accurate even when some of its training modalities are missing at test-time.

Model scope. We employ SegFormer (Xie et al., 2021) as the concrete Transformer in this work, since our focus is semantic segmentation and SegFormer is a popular, efficient, and high-performing choice for this task. However, nothing about our method is specific to SegFormer, and we confirm in Section 4.5 that it transfers well to another popular Transformer: SwinV2 (Liu et al., 2022).

3.1 Overview of A-MoM

A single Transformer with shared weights processes all m modalities in parallel by stacking them along the batch dimension (Figure 2). We insert a lightweight fusion method—A-MoM—between consecutive Transformer layers within all stages. Let $\boldsymbol{x}_i \in \mathbb{R}^d$ be a token of modality i at the output of a given layer, and $\mathbb{X} = \{\boldsymbol{x}_j\}_{j \in \mathbb{M}}$ be the corresponding tokens of all modalities. The input token to the next layer for modality i is

$$f_i(\mathbb{X}) \;=\; \sum_{j \in \mathbb{M}} \underbrace{s_i(\boldsymbol{x}_i)_j}_{\text{scorer}} \cdot \underbrace{a_{j \to i}(\boldsymbol{x}_j)}_{\text{adapter}}, \quad s_i(\boldsymbol{x}_i) = \text{softmax}(\boldsymbol{W}_i \boldsymbol{x}_i + \boldsymbol{b}_i)\,, \quad \boldsymbol{W}_i \in \mathbb{R}^{m \times d}, \boldsymbol{b}_i \in \mathbb{R}^m, \tag{1}$$

where the scorer produces one weight per modality and the adapters map the source token $\boldsymbol{x}_j$ into the feature space of the target modality i. Equation (1) runs in parallel over all tokens and all target modalities i. The computational cost of this is low, as we analyze in Section 3.6 and measure in Section 4.7. At every stage output, a learned aggregation collapses the m per-modality tokens into the single token the prediction head expects. The subsequent sections expand on this overview.

3.2 Token-level Scoring

Motivation. Different modalities excel in different areas of a scene (Figure 1): an RGB camera can read a road marking that a depth sensor simply cannot see. An event camera has high dynamic range, enabling it to perceive the tunnel exit. Depth makes the cyclist stand out clearly. These are but a few examples of how sensors have local advantages. This motivates a local context-dependent approach to fusion. Since the token is the smallest (most local) unit a Transformer operates on, we propose a token-level fusion function. This is the role of the scorer in Equation (1).

Final aggregation. We have m tokens per position, one per modality, whereas the prediction head needs one. We resolve this with a similar mechanism and apply it at every stage output: a modality-specific scorer that learns how much to attend to each modality's final representation,

$$\boldsymbol{x}_{\text{final}} = \sum_{i \in \mathbb{M}} \alpha_i \cdot \boldsymbol{x}_i, \qquad \boldsymbol{\alpha} = \text{softmax}\Big(\boldsymbol{W}_1^f \boldsymbol{x}_1 + b_1^f,\; \ldots,\; \boldsymbol{W}_m^f \boldsymbol{x}_m + b_m^f\Big)\,, \tag{2}$$

with $\boldsymbol{W}_i^f \in \mathbb{R}^{1 \times d}$ and $b_i^f \in \mathbb{R}$ learnable. We initialize $\boldsymbol{W}_i^f = \boldsymbol{0}$ and $b_i^f = 0$ so that all modalities start out equally important. This step is necessary for prediction heads that are not natively multimodal, such as the SegFormer head. However, it is not part of our core contribution and we use a learned weighted sum for consistency with Equation (1).

3.3 SCORER INITIALIZATION

We want $s_i(\boldsymbol{x}_i)$ to start out concentrated on modality i itself, because we hypothesize that high levels of mixing early in training make every modality come to rely on the others being present, leading to poor missing-modality robustness at test-time. Standard initialization does not give us this.

Proposition 1. *Under Xavier (Glorot & Bengio, 2010) and Kaiming (He et al., 2015) initialization,* $\mathbb{E}[s_i(\boldsymbol{x}_i)_j] = 1/m$ *for all* $j \in \mathbb{M}$.

Proof. We briefly outline the proof here. The entries of $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ are i.i.d., leading to exchangeable scores for a fixed $\boldsymbol{x}_i$. Softmax preserves exchangeability and always sums to 1. It follows that $\mathbb{E}[s_i(\boldsymbol{x}_i)_j] = 1/m$ for $j \in \{1, \ldots, m\}$. See full proof in Appendix A.

We instead initialize the scorer so that a chosen fraction p of the weight sits on the target modality.

Proposition 2. *Let* $p \in (0, 1)$ *and initialize* $\boldsymbol{W}_i = \boldsymbol{0}$, $b_{ii} = 0$, *and*

$$b_{ij} = \log\left(\frac{1-p}{p \cdot (m-1)}\right) \quad \textit{for} \quad i \neq j. \tag{3}$$

Then $s_i(\boldsymbol{x}_i)_i = p$ *and* $s_i(\boldsymbol{x}_i)_j = (1-p)/(m-1)$ *for* $i \neq j$.

Proof. Given that $\boldsymbol{W}_i = 0$, we have $\boldsymbol{W}_i\boldsymbol{x}_i + \boldsymbol{b}_i = \boldsymbol{b}_i$. Therefore, we only need to prove that $\mathrm{softmax}(b_i)_i = p$, from which it follows that $\mathrm{softmax}(\boldsymbol{b}_i)_j = (1-p)/(m-1)$ for $i \neq j$ since there is $1-p$ probability mass remaining and b_{ij} is initialized to the same value, so the remaining mass must be distributed equally among them. We prove $\mathrm{softmax}(\boldsymbol{b}_i)_i = p$ in Appendix B.

What *p* controls. $f_i(\mathbb{X}) \to \boldsymbol{x}_i$ as $p \to 1$ at initialization (since $a_{i\to i}(\boldsymbol{x}) = \boldsymbol{x}$), effectively preventing any fusion. At the other end, $p = 1/m$ recovers the uniform initialization of Proposition 1 that leads to high mixing. As such, p is a knob to provide control between these extremes at initialization. In our experiments (see Section 4.6), we find that $p = 0.9$ works well.

3.4 ADAPTERS BETWEEN PAIRS OF MODALITIES

We argue in Section 3.2 that a modality's advantage is local. A second, distinct kind of difference concerns the representation itself: RGB yields dense textures, an event camera yields spatially sparse brightness changes, and LiDAR yields sparse 3D geometry. The representations live in different spaces, so it stands to reason that modality j should be mapped to modality i's feature space prior to mixing. The target modality i already resides in its own feature space and needs no mapping:

$$a_{j\to i}(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} & i = j \\ \mathrm{MLP}_{j\to i}(\boldsymbol{x}) & i \neq j. \end{cases} \tag{4}$$

Writing $g_{d_{\text{in}}\to d_{\text{out}}}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ for a generic Kaiming-initialized (He et al., 2015) linear layer with $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ and $\boldsymbol{b} \in \mathbb{R}^{d_{\text{out}}}$, we define each adapter as a down-projection to a bottleneck of rank r, a GELU nonlinearity, and an up-projection back to d:

$$\mathrm{MLP}_{j\to i} = g_{r\to d} \circ \mathrm{GELU} \circ g_{d\to r}, \tag{5}$$

with $r = 8$ in line with prior work (Li et al., 2025a).

3.5 INFERENCE WITH MISSING MODALITIES

A-MoM is trained once on all of the modalities $\mathbb{M}$ and requires no change of weights to run on a non-empty subset $\mathbb{M}_{miss}$. We simply restrict both softmaxes to the modalities that are available, effectively masking the logits of absent modalities to $-\infty$ before normalizing:

$$f_i(\mathbb{X}) = \sum_{j\in\mathbb{M}_{miss}} \tilde{s}_i(\boldsymbol{x}_i)_j \cdot a_{j\to i}(\boldsymbol{x}_j), \qquad \tilde{s}_i(\boldsymbol{x}_i)_j = \frac{\exp\left([\boldsymbol{W}_i\boldsymbol{x}_i + \boldsymbol{b}_i]_j\right)}{\sum_{k\in\mathbb{M}_{miss}} \exp\left([\boldsymbol{W}_i\boldsymbol{x}_i + \boldsymbol{b}_i]_k\right)}, \tag{6}$$

for $i \in \mathbb{M}_{miss}$, and identically for $\boldsymbol{\alpha}$ in Equation (2). Importantly, renormalizing keeps the scale of f_i comparable to the scale seen during training. Regarding efficiency, absent modalities are dropped from the batch entirely rather than zero-filled, so their tokens are never computed and inference is correspondingly cheaper for smaller $\mathbb{M}_{miss}$.

Table 1: **Robustness scores.** A-MoM matches or exceeds the baselines on all datasets, resulting in an average robustness score improvement of 1.5 over the next-best baseline. The results confirm that A-MoM is robust to missing modalities. **Best** and next-best.

	#Params (M)	DSEC	FMB	MCubeS	MUSES	Avg
CHARM	43.6	49.1	53.5	32.8	47.6	45.7
CMNeXt	58.1	38.8	48.5	24.7	29.6	35.4
GeminiFusion	40.7	43.0	51.5	29.5	41.3	41.3
MAGIC	28.4	50.1	54.4	34.6	50.0	47.3
StitchFusion	26.6	49.2	54.4	40.9	51.9	49.1
TokenFusion	35.5	46.2	54.1	32.3	32.4	41.3
A-MoM (ours)	26.6	**50.5**	**54.9**	**42.6**	**54.3**	**50.6**

Table 2: **Weather and lighting conditions.** Breakdown of overall mIoU on MUSES. A-MoM proves robust to adverse conditions, notably raising performance in fog by 7.6. **Best** and next-best.

	Clear	Fog	Rain	Snow	Day	Night
CHARM	57.2	62.7	58.2	58.4	64.2	63.4
CMNeXt	53.7	48.3	51.0	56.3	58.6	56.4
GeminiFusion	63.8	67.3	64.3	66.7	**73.0**	67.6
MAGIC	62.5	58.7	61.5	65.7	64.5	68.2
StitchFusion	64.0	66.9	**64.5**	67.0	70.9	69.8
TokenFusion	50.1	55.4	46.5	48.7	55.1	50.8
A-MoM (ours)	**64.2**	**74.9**	64.2	**68.2**	72.7	**69.9**

3.6 Efficiency

A-MoM adds $\mathcal{O}(m^2 \cdot d \cdot r)$ parameters and $\mathcal{O}(n \cdot m^2 \cdot d \cdot r)$ time per layer, for n tokens; Appendix C derives both. The quadratic influence of m is small in practice because m is small ($m \in \{2, 3, 4\}$ in our experiments). Likewise, the contribution of r is small as r is intentionally a small bottleneck and we use $r = 8$ in our experiments. Concretely, for $m = 4$ on MiT-B2 (Xie et al., 2021), A-MoM adds 0.8M parameters, or 3.3% of the backbone, and the added components are fully parallel across modalities and tokens. Section 4.7 measures the resulting wall-clock cost.

4 Experiments

4.1 Setup

Baselines. We compare to 6 representative and strong baselines. The baselines are chosen to enable a solid evaluation of A-MoM, both in terms of the robustness to missing modalities and the ideal case with all modalities being present: CHARM (Wen et al., 2026), CMNeXt (Zhang et al., 2023), and MAGIC (Zheng et al., 2025b) are designed for missing-modality robustness and serve as representative works in this domain. TokenFusion (Wang et al., 2022) is a defining work for exchange-based models, like A-MoM, where GeminiFusion (Jia et al., 2024) and StitchFusion (Li et al., 2025a) stand as high-performing methods within this paradigm.

Datasets. We evaluate A-MoM on 4 datasets spanning 8 diverse modalities, showcasing its broad applicability. DSEC (Gehrig et al., 2021; Sun et al., 2022) has 11K automotive samples with **RGB** and **events** modalities, annotated with 11 categories. In addition, we derive **depth** from the stereo RGB cameras using FoundationStereo (Wen et al., 2025), totaling 3 modalities. FMB (Liu et al., 2023) consists of 1.5K automotive **RGB** and infrared (**IR**) samples across 14 categories. MCubeS (Liang et al., 2022) deals with material segmentation and has 500 samples of 20 material categories with 4 modalities: **RGB**, near-infrared (**NIR**), angle of linear polarization (**AoLP**), and degree of linear polarization (**DoLP**). MUSES (Brödermann et al., 2025a) provides 2.5K automotive samples with **RGB**, **events**, and **LiDAR** across 19 categories and under diverse weather and lighting conditions.

Table 3: **Robustness vs specialization.** Ratios (as percentages) of mIoU on MCubeS for the model trained on all modalities, $\mathbb{M}$, and the model trained on the specific modality-subset, $\mathbb{M}_{miss}$, specified by the column. Legend: **A**oLP, **D**oLP, **N**ear-infrared, **R**GB. **Best** and next-best.

	A	D	N	R	A-D	A-N	D-N	A-R	D-R	N-R	A-D-N	A-D-R	A-N-R	D-N-R	Avg
CHARM	53.3	37.9	49.2	47.2	70.6	70.9	76.8	83.3	81.8	84.4	88.6	97.4	96.4	92.9	73.6
CMNeXt	12.8	7.4	3.0	80.5	12.4	11.9	11.3	83.3	**97.8**	78.1	27.4	93.1	91.1	95.1	50.4
GeminiFusion	12.5	11.7	4.3	93.3	20.8	36.5	21.0	91.0	97.1	**97.2**	60.8	98.6	96.8	100.5	60.2
MAGIC	38.3	47.9	49.1	83.9	56.6	63.0	67.6	87.0	88.6	88.5	71.7	90.0	88.8	93.0	72.4
StitchFusion	59.3	83.3	80.2	**93.6**	80.3	84.4	90.0	93.8	93.3	90.6	92.3	97.2	95.6	96.7	87.9
TokenFusion	29.7	51.4	46.6	81.3	56.4	66.0	76.0	88.1	94.5	89.2	82.4	94.7	90.2	96.8	74.5
A-MoM (ours)	**72.0**	**85.1**	**85.5**	93.4	**90.1**	**90.4**	**94.6**	**94.8**	97.6	96.0	**97.0**	**100.7**	**97.9**	**103.0**	**92.7**

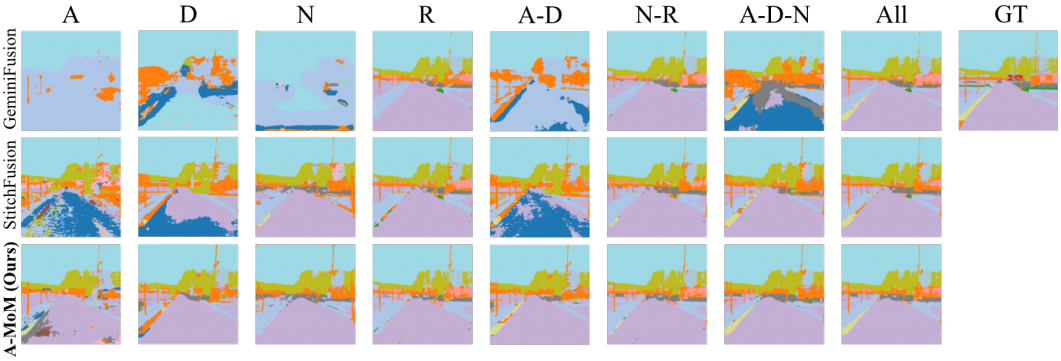

Figure 3: **Qualitative results.** Predictions of GeminiFusion, StitchFusion, and A-MoM on 8 modality-subsets on MCubeS. A-MoM yields meaningful predictions even in the extreme cases of only a single modality being available. Legend: **A**oLP, **D**oLP, **N**ear-infrared, **R**GB, **G**round **T**ruth.

Since events and LiDAR are not natively represented as 2D images, we need to transform them into 2D representations: events become a Stacked Histogram (Gehrig et al., 2019) from the most recent 50 ms with 5 time buckets, and we project LiDAR points into the 2D view using intrinsics and extrinsics.

Training. We implement all models in PyTorch (Ansel et al., 2024) using 16-bit mixed precision and optimize them with AdamW (Loshchilov & Hutter, 2019). All models are initialized from the official SegFormer (Xie et al., 2021) checkpoints using B2 as the default size. Crucially, we fix all common hyperparameters across fusion methods. The only systematic difference between models is the choice of fusion method. The setup is based on Zhang et al. (2023) and is not tuned to A-MoM. This setup yields a fair apples-to-apples comparison of the fusion methods. See Appendix F for details.

Evaluation. The average performance is the primary interest of this work, as it reflects the general robustness to missing modalities. Therefore, we report the average mIoU over all non-empty subsets of the training modalities and refer to it as the *robustness score*. This reflects how the model performs in both ideal and non-ideal cases. We also report the *overall* mIoU as the performance obtained in the ideal setting where all training modalities are available at test-time.

4.2 Robustness 2

Robustness to missing modalities. Table 1 reports the robustness scores. A-MoM matches or exceeds the baselines on all 4 datasets. A-MoM also raises the average robustness score by 1.5 over the next-best baseline, StitchFusion. These results highlight that A-MoM is robust to missing modalities. We break down the performance on all subsets and datasets in Appendix E.

Adverse weather and lighting conditions. Beyond hard failures, sensors may degrade under adverse conditions such as poor visibility for RGB. Table 2 reports the performance on MUSES broken down by weather and lighting conditions. A-MoM is the highest-performing method on 4 of 6 conditions, closely trailing the best baseline by 0.3 in the remaining 2 cases. A-MoM is particularly robust in fog and snow with respective increases of at least 7.6 and 1.2 over the baselines. We find that A-MoM is robust to adverse weather and lighting conditions.

Table 4: **Overall scores.** A-MoM is overall a strong fusion method. It ranks highest on 2 datasets and next-highest on another 2 datasets, increasing the average overall score by 0.7. A-MoM ranks highest on 3 of 4 datasets when reducing the backbone size of GeminiFusion (B2 $\rightarrow$ B1) to make parameter budgets more comparable. Legend: **best**, <u>next-best</u>, and † MiT-B1 backbone size.

	Size	#Params (M)	DSEC	FMB	MCubeS	MUSES	Avg
CHARM	B2	43.6	66.2	59.5	44.9	64.7	58.8
CMNeXt	B2	58.1	66.2	59.3	47.7	59.6	58.2
GeminiFusion†	B1	22.5	64.1	60.6	45.9	64.9	58.9
GeminiFusion	B2	40.7	66.6	59.6	**50.8**	<u>71.8</u>	62.2
MAGIC	B2	28.4	65.9	63.7	45.2	66.0	60.2
StitchFusion	B2	26.6	<u>67.3</u>	**64.5**	48.1	71.6	<u>62.9</u>
TokenFusion	B2	35.5	61.4	63.3	40.8	55.1	55.1
A-MoM (ours)	B2	26.6	**67.5**	<u>64.4</u>	<u>49.6</u>	**72.9**	**63.6**

Robustness vs specialization. We have established A-MoM to be robust. However, this does not answer how close it is to its full potential. To be perfectly robust, a model trained on $\mathbb{M}$ but evaluated on $\mathbb{M}_{miss} \subset \mathbb{M}$ should match the performance of a specialized model trained on $\mathbb{M}_{miss}$ and evaluated on $\mathbb{M}_{miss}$. Table 3 reports the ratios of the scores from models trained on the full MCubeS and each of the 14 modality-subsets, excluding the full set of modalities. We train a total of $7 \times 14 = 98$ specialist models for this experiment. The ratios inform how close the model is to obtaining its ideal robustness. The average ratio of A-MoM is high at 92.7% with StitchFusion being next-best at 87.9%, in line with the findings from Table 1. A-MoM narrows the specialization gap, bringing robust multimodal models closer to their full potential.

Qualitative results. Figure 3 shows predictions of GeminiFusion, StitchFusion, and A-MoM for several modality-subsets on MCubeS. StitchFusion struggles when RGB and near-infrared are missing, and GeminiFusion heavily depends on RGB—leading to highly unreliable predictions without it. A-MoM is robust in all cases.

4.3 Overall Performance 3

While missing-modality robustness is the focus of this work, we also care about the performance in the ideal case where all modalities are available at test-time, and we report this in Table 4. In this setting, A-MoM remains a strong fusion method as it is the highest-ranking on DSEC ($+0.2$) and MUSES ($+1.1$) and next-highest on FMB (-0.1) and MCubeS (-1.2). The average overall score improvement is 0.7 over the next-best, StitchFusion. In addition, we see that GeminiFusion comes out as the highest-ranking on MCubeS and next-highest on MUSES, but Table 1 reports that GeminiFusion underperforms on robustness, with differences in robustness scores to A-MoM ranging from 3.4 to 13.1. As such, this highlights that only measuring the overall case is insufficient for understanding the performance of a model under real-world conditions.

Parameter budgets. GeminiFusion wins on MCubeS in Table 4. Yet, it increases the parameter count over A-MoM by 14.1M, corresponding to a 53% increase in parameters. This raises the question whether its overall scores are due to stronger fusion or because it increases the model size. To answer this, we reduce its parameter budget by reducing the backbone size from MiT-B2 to MiT-B1, thus obtaining GeminiFusion†. While this is not an exact parameter match, it is the closest possible match within the MiT family of models. The parameter difference to A-MoM becomes -4.1M after the reduction, leading to substantial drops in overall scores, with A-MoM consistently outperforming GeminiFusion†. As such, we attribute the differences on MCubeS in Table 4 to larger models and not to GeminiFusion being a stronger fusion method in the overall case.

4.4 Scaling Study

The results so far focus on a single scale (MiT-B2), raising the question of how A-MoM scales. To answer this question, we compare A-MoM and the baselines on DSEC using 5 scales (MiT-B0–B4). Figure 4 shows the results. A-MoM scales well as both the robustness and overall performance are

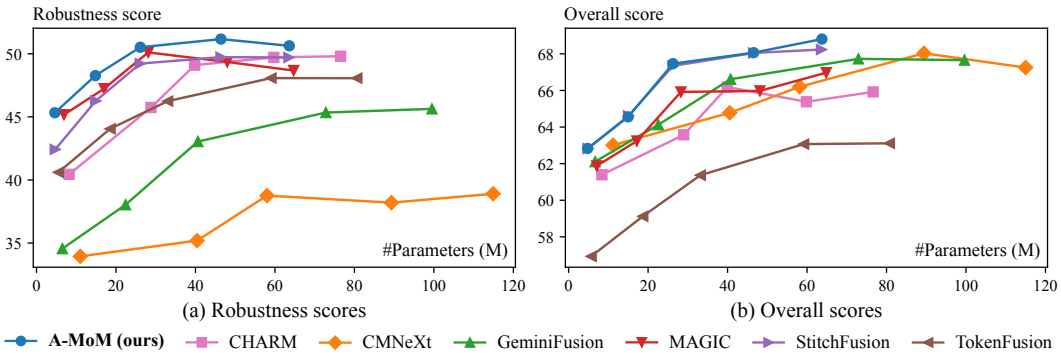

Figure 4: **Scaling study on DSEC.** A-MoM preserves its advantage at 5 different scales (MiT-B0–B4) for both the (a) robustness and (b) overall cases.

Table 5: **SwinV2 robustness scores.** A-MoM remains robust to missing modalities, showing that A-MoM is not only applicable to MiT. **Best** and next-best.

	#Params (M)	DSEC	FMB	MCubeS	MUSES	Avg
GeminiFusion	90.2	41.2	49.4	29.1	40.2	40.0
StitchFusion	52.6	49.5	**56.6**	40.9	54.4	50.4
A-MoM (ours)	52.6	**50.3**	56.2	**41.9**	**55.6**	**51.0**

consistently high for all scales, highlighting that A-MoM is scalable for the considered model sizes, and setting a Pareto front in terms of performance vs. parameters.

4.5 Model Study

We have chosen SegFormer/MiT (Xie et al., 2021) as the underlying model in this work. However, A-MoM is not inherently tied to MiT and here we assess how well it generalizes to another popular model: SwinV2 (Liu et al., 2022). We focus this experiment on the strongest baselines: GeminiFusion and StitchFusion. Table 5 reports the robustness scores using the small variant of SwinV2. Compared to Table 1, the gap between StitchFusion and A-MoM is smaller and StitchFusion is slightly more robust on FMB, but the general finding remains: A-MoM shows high robustness and is the most robust on 3 of 4 datasets. As such, the results are well-aligned with those observed for MiT in Table 1, confirming that A-MoM generalizes to SwinV2 in the robustness case; Appendix D shows the same holds for the overall case. We find that A-MoM's performance is not tied specifically to MiT.

Table 6: **Ablation study** on the DSEC validation split. (a) Ablating design choices confirms their usefulness. (b) Ablating p finds $p = 0.9$ to be the best choice. **Best** and next-best.

(a)

Ablation	Robustness	Overall
Full (Eq. 1 and 2)	51.9	67.5
- Eq. 2 only	48.0 (−3.9)	65.2 (−2.3)
- Uniform scorer	39.2 (−12.7)	63.1 (−4.4)
- Shared scorers	50.0 (−1.9)	67.1 (−0.4)
- Kaiming init	39.0 (−12.9)	65.8 (−1.7)
- No adapters	50.2 (−1.7)	65.8 (−1.7)

(b)

p	Robustness	Overall
0.33 ($1/m$)	43.1	65.5
0.50	46.0	66.5
0.66	48.3	67.0
0.75	49.2	67.1
0.90	**51.9**	**67.5**
0.99	50.4	66.9

Table 7: **Inference time (ms).** Measured on an RTX 4090 using MiT-B2 with a batch size of 1. A-MoM is overall fast, ranking next-fastest on average with the average inference time being only 1.5 ms behind the fastest baseline. **Fastest** and next-fastest.

Model	DSEC	FMB	MCubeS	MUSES	Avg
CHARM	15.9	17.2	17.6	25.6	19.1
CMNeXt	**14.3**	17.4	**14.4**	**19.3**	**16.4**
GeminiFusion	19.7	29.1	17.4	32.2	24.6
MAGIC	16.6	19.1	17.6	30.1	20.9
StitchFusion	24.4	19.0	35.3	32.6	27.8
TokenFusion	20.0	20.8	21.4	29.3	22.9
A-MoM (ours)	14.8	**16.7**	15.7	24.5	17.9

4.6 Ablation Study

We systematically ablate each component to understand its contribution. Table 6a reports this on the DSEC validation split. A-MoM (Equation (1)) drives much of the performance as only relying on Equation (2) leads to a robustness drop of 3.9 and an overall drop of 2.3. We ablate the learned scorers by averaging the modalities, corresponding to a uniform distribution. This greatly diminishes the robustness (-12.7) and overall (-4.4) performance, confirming that learned scorers contribute significantly. Likewise, each modality must learn its own scorer to preserve robustness, since sharing leads to a 1.9 drop in robustness. Our initialization of $\boldsymbol{W}_i$, $\boldsymbol{b}_i$, $\boldsymbol{W}_i^f$, and b_i^f contributes greatly to A-MoM's performance as we observe a large drop in robustness (-12.9) and overall (-1.7) performance when using Kaiming initialization. Table 6b further studies the choice of p, finding that $p = 0.9$ obtains the highest robustness and overall scores. Finally, we confirm that adapters contribute meaningfully as removing them causes a 1.7 drop in both robustness and overall performance.

4.7 Inference Time

We report inference time in Table 7, measured on an RTX 4090 using MiT-B2 and a batch size of 1. Overall, we find A-MoM to be fast as it ranks next-fastest on average with CMNeXt being the fastest. We note that CMNeXt merges all non-RGB modalities into a single stream, effectively yielding near-identical inference times for any $m \geq 2$. This results in low inference times for $m > 2$ as we see for DSEC ($m = 3$), MCubeS ($m = 4$), and MUSES ($m = 3$). Nevertheless, A-MoM is highly parallelizable with only a slight overhead from its contributed components, resulting in an efficient method with a small increase in average inference time of 1.5 ms compared to CMNeXt.

5 Limitations & Future Work

We find A-MoM to be a robust fusion method for multimodal semantic segmentation. The experiments confirm this finding for SegFormer and SwinV2. Yet, it is not clear how well A-MoM generalizes beyond this task and these models. We expect it to generalize well, given that A-MoM does not make task- or model-specific design choices, and we plan to validate this in future work. Specifically, we seek to explore robot perception tasks such as object detection and depth estimation.

6 Conclusion

We introduce Adapted Mixture of Modalities (A-MoM), a fusion method for multimodal semantic segmentation that is robust to missing modalities. A-MoM dynamically fuses any number of modalities based on their local context. Extensive experiments across 4 datasets spanning 8 diverse modalities confirm that A-MoM is highly robust to missing modalities (+1.5 robustness score) while remaining a strong method in the ideal case (+0.7 overall score) where all training modalities are available at test-time. Thus, A-MoM makes multimodal semantic segmentation more robust to missing modalities, narrowing the gap to real-world deployment.

AI USE STATEMENT

In this work, we used generative AI tools to *assist in writing proofs*.

We have not used generative AI tools to *help develop theoretical models or conceptual frameworks*, *formulate mathematical claims*, *provide critical ingredients for proving mathematical claims*, *propose or refine hypotheses*, *design or provide feedback on research methodology or experiments*, *implement methods*, or *interpret results*.

The following are not applicable to this work: *generate synthetic datasets*, *assist with translation*, *clean and reformat dataset*, and *support qualitative and thematic data analysis*.

Additionally, we used generative AI tools to *create or edit software code* and *draft parts of a research paper*, and *edit a research paper to improve readability*.

We have reviewed all AI-assisted work. We manually checked all outputs for correctness and never copied anything verbatim. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will make the code and model weights publicly available upon acceptance. The code includes everything needed to run all of the experiments, along with instructions for how to download the datasets and pretrained backbone weights and the exact training config files that we used ourselves. We have further locked package versions, ensuring an easily reproducible setup. Additionally, we detail the training setup in Appendix F. We also detail the proofs of Propositions 1 and 2 in Appendices A and B, respectively.

REFERENCES

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM international conference on architectural support for programming languages and operating systems, volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366. URL https://pytorch.org/assets/pytorch2-2.pdf.

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 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 21–38, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-73202-7. doi: 10.1007/978-3-031-73202-7_2.

Tim Brödermann, Christos Sakaridis, Yuqian Fu, and Luc Van Gool. CAFuser: Condition-Aware Multimodal Fusion for Robust Semantic Perception of Driving Scenes. *IEEE Robotics and Automation Letters*, 10(4):3134–3141, April 2025b. ISSN 2377-3766. doi: 10.1109/LRA.2025.3536218. URL https://ieeexplore.ieee.org/abstract/document/10858375.

Tim Brödermann, Christos Sakaridis, Luigi Piccinelli, Wim Abbeloos, and Luc Van Gool. DGFusion: Depth-Guided Sensor Fusion for Robust Semantic Perception. *IEEE Robotics and Automation Letters*, 11(3):3334–3341, March 2026. ISSN 2377-3766. doi: 10.1109/LRA.2026.3656789. URL https://ieeexplore.ieee.org/abstract/document/11361097.

Jiaxin Cai, Jingze Su, Qi Li, Wenjie Yang, Shu Wang, Tiesong Zhao, Shengfeng He, and Wenxi Liu. Keep the Balance: A Parameter-Efficient Symmetrical Framework for RGB+X Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.

10587–10598, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cai_Keep_the_Balance_A_Parameter-Efficient_Symmetrical_Framework_for_RGBX_Semantic_CVPR_2025_paper.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-End Learning of Representations for Asynchronous Event-Based Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5633–5643, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Gehrig_End-to-End_Learning_of_Representations_for_Asynchronous_Event-Based_Data_ICCV_2019_paper.html.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A Stereo Event Camera Dataset for Driving Scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, July 2021. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2021.3068942. URL https://ieeexplore.ieee.org/document/9387069/.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, March 2010. URL https://proceedings.mlr.press/v9/glorot10a.html.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. FuseMoE: Mixture-of-Experts Transformers for Fleximodal Fusion. *Advances in Neural Information Processing Systems*, 37:67850–67900, December 2024. doi: 10.52202/079017-2167. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7d62a85ebfed2f680eb5544beae93191-Abstract-Conference.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html.

Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. GeminiFusion: Efficient Pixel-wise Multimodal Fusion for Vision Transformer. In *Forty-first International Conference on Machine Learning*, June 2024. URL https://openreview.net/forum?id=Zsz9Pdfvtg.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal Prompting With Missing Modalities for Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Lee_Multimodal_Prompting_With_Missing_Modalities_for_Visual_Recognition_CVPR_2023_paper.html.

Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. StitchFusion: Weaving Any Visual Modalities to Enhance Multimodal Semantic Segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 1308–1317, New York, NY, USA, October 2025a. Association for Computing Machinery. ISBN 979-8-4007-2035-2. doi: 10.1145/3746027.3755110. URL https://dl.acm.org/doi/10.1145/3746027.3755110.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A Simple Framework for Multi-Modal Learning with Missing Modality. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24068–24077, 2025b. doi: 10.1109/ICCV51701.2025.02231. URL https://ieeexplore.ieee.org/abstract/document/11443373.

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*, pp. 19800–19808, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liang_Multimodal_Material_Segmentation_CVPR_2022_paper.html.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8115–8124, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Liu_Multi-interactive_Feature_Learning_and_a_Full-time_Multi-modality_Benchmark_for_Image_ICCV_2023_paper.html.

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)*, pp. 961–968, June 2024. doi: 10.1109/IV55156.2024.10588722. URL https://ieeexplore.ieee.org/abstract/document/10588722.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12009–12019, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Swin_Transformer_V2_Scaling_Up_Capacity_and_Resolution_CVPR_2022_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are Multimodal Transformers Robust to Missing Modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Ma_Are_Multimodal_Transformers_Robust_to_Missing_Modality_CVPR_2022_paper.html.

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*, pp. 1020–1030, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Maheshwari_Missing_Modality_Robustness_in_Semi-Supervised_Multi-Modal_Semantic_Segmentation_WACV_2024_paper.html.

Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif. Robust Multimodal Learning With Missing Modalities via Parameter-Efficient Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):742–754, February 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3476487. URL https://ieeexplore.ieee.org/abstract/document/10713849.

Leslie N. Smith and Nicholay Topin. Super-Convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pp. 369–386. SPIE, May 2019. doi: 10.1117/12.2520589. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11006/1100612/Super-convergence--very-fast-training-of-neural-networks-using/10.1117/12.2520589.full.

Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. ESS: Learning Event-Based Semantic Segmentation from Still Images. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pp. 341–357, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19830-4. doi: 10.1007/978-3-031-19830-4_20.

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*, pp. 12186–12195, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Multimodal_Token_Fusion_for_Vision_Transformers_CVPR_2022_paper.html.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-Shot Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5249–5260, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wen_FoundationStereo_Zero-Shot_Stereo_Matching_CVPR_2025_paper.html.

Lekang Wen, Jing Xiao, Liang Liao, Jiajun Chen, and Mi Wang. CHARM: Collaborative Harmonization Across Arbitrary Modalities for Modality-Agnostic Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(13):10603–10611, March 2026. ISSN 2374-3468. doi: 10.1609/aaai.v40i13.38033. URL https://ojs.aaai.org/index.php/AAAI/article/view/38033.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

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*, pp. 1136–1147, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Delivering_Arbitrary-Modal_Semantic_Segmentation_CVPR_2023_paper.html.

Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105, January 2021. ISSN 0262-8856. doi: 10.1016/j.imavis.2020.104042. URL https://www.sciencedirect.com/science/article/pii/S0262885620301748.

Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning Modality-Agnostic Representation for Semantic Segmentation from Any Modalities. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 146–165, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-72754-2. doi: 10.1007/978-3-031-72754-2_9.

Xu Zheng, Yuanhuiyi Lyu, Jiazhou Zhou, and Lin Wang. Centering the Value of Every Modality: Towards Efficient and Resilient Modality-Agnostic Semantic Segmentation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 192–212, Cham, 2025b. Springer Nature Switzerland. ISBN 978-3-031-72890-7. doi: 10.1007/978-3-031-72890-7_12.

APPENDIX

A PROOF OF PROPOSITION 1

We recall that Proposition 1 states that standard initialization, such as Xavier (Glorot & Bengio, 2010) and Kaiming (He et al., 2015) initialization, makes $s_i(\boldsymbol{x}_i)$ yield uniform scores in expectation:

$$\mathbb{E}[s_i(\boldsymbol{x}_i)_j] = \frac{1}{m} \quad \text{for} \quad j \in \{1, \dots, m\}. \tag{7}$$

Likewise, recall that

$$s_i(\boldsymbol{x}_i) = \operatorname{softmax}(\boldsymbol{W}_i \boldsymbol{x}_i + \boldsymbol{b}_i) \quad \text{with} \quad \boldsymbol{W}_i \in \mathbb{R}^{m \times d}, \boldsymbol{b}_i \in \mathbb{R}^m. \tag{8}$$

Let $z_j = \boldsymbol{w}_j^T \boldsymbol{x}_i + b_{ij}$ where $\boldsymbol{w}_j$ is the jth row of $\boldsymbol{W}_i$. The proof of Equation (7) follows 3 steps: (1) $\boldsymbol{z}$ is exchangeable, (2) softmax preserves exchangeability, and (3) softmax of exchangeable scores is uniform in expectation.

1) $\boldsymbol{z}$ is exchangeable. The initialization of $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ draws each entry i.i.d., so the pairs $(\boldsymbol{w}_1, b_{i1}), \ldots, (\boldsymbol{w}_m, b_{im})$ are also i.i.d. Since $\boldsymbol{x}_i$ is fixed, it follows that $z_1, \ldots, z_m$ are i.i.d. This lets us see that for any permutation π, $\boldsymbol{z}$ and the permuted sequence are equal in distribution

$$(z_1, \ldots, z_m) \overset{d}{=} (z_{\pi(1)}, \ldots, z_{\pi(m)}), \tag{9}$$

which satisfies the definition of being exchangeable.

2) Softmax preserves exchangeability. Softmax applies the exponential function to each element of a vector and divides by a common denominator, $\sum_j \exp z_j$, which is permutation-invariant because summation commutes. We therefore get that $\mathrm{softmax}(\pi(\boldsymbol{z})) = \pi(\mathrm{softmax}(\boldsymbol{z}))$ for any permutation π. Combining this with step 1 gives exchangeability for s_i:

$$(s_i(\boldsymbol{x}_i)_1, \ldots, s_i(\boldsymbol{x}_i)_m) \overset{d}{=} (s_i(\boldsymbol{x}_i)_{\pi(1)}, \ldots, s_i(\boldsymbol{x}_i)_{\pi(m)}) \tag{10}$$

3) Softmax of exchangeable scores is uniform in expectation. We know that softmax by design sums to 1, so the expectation is also 1:

$$\mathbb{E}\left[\sum_j s_i(\boldsymbol{x}_i)_j\right] = \sum_j \mathbb{E}\left[s_i(\boldsymbol{x}_i)_j\right] = 1. \tag{11}$$

The exchangeability of s_i implies that

$$\mathbb{E}\left[s_i(\boldsymbol{x}_i)_j\right] = \mathbb{E}\left[s_i(\boldsymbol{x}_i)_k\right] \tag{12}$$

holds for all $j, k \in \{1, \ldots, m\}$, allowing us to write Equation (11) as

$$\sum_k \mathbb{E}\left[s_i(\boldsymbol{x}_i)_k\right] = m \cdot \mathbb{E}\left[s_i(\boldsymbol{x}_i)_j\right] = 1 \quad \text{for} \quad j \in \{1, \ldots, m\}. \tag{13}$$

Rearranging the terms completes the proof:

$$\mathbb{E}\left[s_i(\boldsymbol{x}_i)_j\right] = \frac{1}{m} \quad \text{for} \quad j \in \{1, \ldots, m\} \quad \square \tag{14}$$

B Proof of Proposition 2

Recall from Section 3.3 that the bias term of each scorer is initialized as

$$b_{ij} = \begin{cases} 0 & i = j \\ \log\left(\frac{1-p}{p\cdot(m-1)}\right) & i \neq j \end{cases}, \tag{15}$$

and we need to prove that $\mathrm{softmax}(\boldsymbol{b}_i)_i = p$ for $p \in (0, 1)$. We prove this by expanding the softmax:

$$\begin{aligned}
\mathrm{softmax}(\boldsymbol{b}_i)_i &= \frac{\exp b_{ii}}{\exp b_{ii} + \sum_{j\neq i} \exp b_{ij}} && (16)\\
&= \frac{1}{1 + \sum_{j\neq i} \exp b_{ij}} \quad \text{using} \quad b_{ii} = 0 && (17)\\
&= \frac{1}{1 + (m-1)\cdot\left(\frac{1-p}{p\cdot(m-1)}\right)} \quad \text{using} \quad b_{ij} = \log\left(\frac{1-p}{p\cdot(m-1)}\right) && (18)\\
&= \frac{1}{1 + \frac{1-p}{p}} && (19)\\
&= \frac{p}{p + (1-p)} \quad \text{multiplying} \quad \frac{p}{p} && (20)\\
&= p \quad \square && (21)
\end{aligned}$$

Table 8: **SwinV2 overall scores.** A-MoM retains its position with the highest average overall score (64.5 vs 63.7) and matches or exceeds the baselines on 3 datasets: DSEC, FMB, and MUSES. The results are consistent with the MiT results in Table 4, confirming that A-MoM is applicable beyond MiT. **Best** and next-best.

	#Params (M)	DSEC	FMB	MCubeS	MUSES	Avg
GeminiFusion	90.2	64.6	62.1	**51.1**	71.6	62.3
StitchFusion	52.6	66.1	66.0	48.9	73.8	63.7
A-MoM (ours)	52.6	**66.2**	**67.5**	50.1	**74.2**	**64.5**

C Efficiency Analysis

We analyze the per-layer time and parameter complexities of A-MoM. For this analysis, n is the number of tokens in a given layer. We assume $d > 1$, which is true for any model of practical interest, and $d > r$, which is true by design.

C.1 Time Complexity

Scorer cost. Each of the m scorers applies a linear function, costing $m \cdot d + m$ operations per token. Summing over scorers and tokens, we get

$$O(n \cdot m \cdot (m \cdot d + m)) = O(n \cdot m^2 \cdot d). \tag{22}$$

Adapter cost. There are $m \cdot (m-1) = O(m^2)$ adapters, not counting the identity mappings as they do not incur a time cost. Each adapter applies a simple MLP, costing $O(d \cdot r + r + r \cdot d + d) = O(d \cdot r)$ operations per token. Summing over pairs and tokens, we get

$$O(n \cdot m^2 \cdot d \cdot r). \tag{23}$$

Weighted sum. We perform $m \cdot d$ multiplications for every modality. Summing over tokens and modalities, we get

$$O(n \cdot m \cdot m \cdot d) = O(n \cdot m^2 \cdot d). \tag{24}$$

Total cost. We obtain the total cost by summing the 3 costs, yielding a per-layer time complexity of

$$O\left(n \cdot m^2 \cdot d + n \cdot m^2 \cdot d \cdot r + n \cdot m^2 \cdot d\right) = O\left(n \cdot m^2 \cdot d \cdot r\right). \tag{25}$$

This is efficient since m and r are small with $m \in \{2, 3, 4\}$ and $r = 8$ in our experiments.

C.2 Parameter Complexity

The parameter complexity analysis follows that of the time complexity analysis. The main difference is that parameters do not grow with the token count nor does the weighted sum add any parameters.

Scorer cost. The m linear scorers contribute $m \cdot (m \cdot d + m) = O(m^2 \cdot d)$ parameters.

Adapter cost. We have $m \cdot (m-1)$ adapter MLPs, each adding $d \cdot r + r + r \cdot d + d = O(d \cdot r)$ parameters. The m identity functions do not add any parameters. In total, the adapters increase the number of parameters per layer by $O(m^2 \cdot d \cdot r)$.

Total cost. Summing the scorer cost and adapter cost gives a per-layer parameter complexity of

$$O(m^2 \cdot d + m^2 \cdot d \cdot r) = O(m^2 \cdot d \cdot r). \tag{26}$$

Similar to the time complexity, this overhead is small because m and r are small for all practical purposes and by design, respectively.

D SwinV2 Overall Results

Table 8 reports the overall scores with SwinV2 (Liu et al., 2022). A-MoM ranks highest on 3 of 4 datasets (DSEC, FMB, MUSES) and next-highest on MCubeS. Its average overall score exceeds the next-highest, StitchFusion, by 0.8. The results confirm that A-MoM is a strong fusion method in the overall case for SwinV2, further highlighting its applicability beyond MiT.

Table 9: **DSEC results on each modality-subset.** Legend: **D**epth, **E**vents, **R**GB. **Best** and next-best.

	D	E	R	D-E	D-R	E-R	All	Avg
CHARM	12.7	35.7	56.0	**46.9**	62.8	63.4	66.2	49.1
CMNeXt	1.4	1.6	56.1	22.8	61.5	61.6	66.2	38.8
GeminiFusion	1.4	15.4	58.9	36.0	63.2	59.9	66.6	43.0
MAGIC	**25.0**	30.3	**61.2**	41.9	63.1	63.3	65.9	50.1
StitchFusion	20.3	33.2	60.3	37.0	**63.5**	63.0	67.3	49.2
TokenFusion	4.4	36.6	55.9	46.4	59.0	60.0	61.4	46.2
A-MoM (ours)	19.6	**37.5**	58.9	43.4	63.3	**63.5**	**67.5**	**50.5**

Table 10: **FMB results on each modality-subset.** Legend: **I**nfrared, **R**GB. **Best** and next-best.

	I	R	All	Avg
CHARM	**45.7**	55.3	59.5	53.5
CMNeXt	33.8	52.5	59.3	48.5
GeminiFusion	41.9	52.9	59.6	51.5
MAGIC	42.5	**57.0**	63.7	54.4
StitchFusion	43.5	55.1	**64.5**	54.4
TokenFusion	43.1	56.0	63.3	54.1
A-MoM (ours)	44.8	55.4	64.4	**54.9**

E Results per Dataset

Section 4.2 reports robustness results averaged over all modality-subsets. Here we report the mIoU of each modality-subset and dataset in Tables 9 to 12.

F Implementation Details

All models are optimized with AdamW (Loshchilov & Hutter, 2019), using a OneCycle (Smith & Topin, 2019) learning rate schedule. Table 13 reports the training hyperparameters.

All modalities are augmented using these geometric transformations during training: random horizontal flip (50% probability) and random scale + crop (scale randomly selected from the $[0.5, 2.0]$ interval). The RGB modality is further augmented with ColorJitter.

When a dataset has a dedicated validation split (DSEC and MCubeS), we save the checkpoint that obtains the highest overall mIoU and use that for evaluation on the test split. For all other datasets we use the final checkpoint after 500 epochs.

The baselines infer with missing modalities by masking them out whenever summing or averaging over them. Masking corresponds to how A-MoM drops missing modalities (Section 3.5) as it ensures the modalities do not contribute to the prediction. Baselines that assume a fixed number of modalities receive zero-filled inputs for the missing ones.

Table 11: **MCubeS results on each modality-subset.** Legend: **A**oLP, **D**oLP, **N**ear-infrared, **R**GB. **Best** and next-best.

	A	D	N	R	A-D	A-N	D-N	A-R	D-R	N-R	A-D-N	A-D-R	A-N-R	D-N-R	All	Avg
CHARM	20.4	15.1	20.5	22.3	29.1	30.3	32.9	37.3	37.1	37.8	37.7	42.8	42.4	41.9	44.9	32.8
CMNeXt	3.9	2.5	1.1	37.8	4.2	4.4	4.3	39.8	45.2	37.2	10.3	44.6	42.9	44.6	47.7	24.7
GeminiFusion	4.7	4.6	1.8	43.7	8.3	15.8	9.0	43.4	45.2	**46.5**	24.5	47.5	**48.1**	**49.1**	**50.8**	29.5
MAGIC	15.0	19.1	21.5	41.1	23.2	28.4	31.0	41.6	43.5	43.3	33.1	43.4	44.0	45.0	45.2	34.6
StitchFusion	23.1	**33.6**	34.3	**44.3**	33.8	38.1	40.0	45.6	46.6	44.0	42.1	47.5	46.1	46.6	48.1	40.9
TokenFusion	11.3	20.1	19.8	39.1	21.7	27.2	31.8	39.5	41.3	40.0	32.8	40.4	37.8	41.3	40.8	32.3
A-MoM (ours)	**27.8**	32.9	**36.4**	43.8	**37.0**	**40.4**	**41.5**	**46.3**	**47.7**	46.0	**43.8**	**48.7**	47.6	48.7	49.6	**42.6**

Table 12: **MUSES results on each modality-subset.** Legend: **E**vents, **L**iDAR, **R**GB. **Best** and next-best.

	E	L	R	E-L	E-R	L-R	All	Avg
CHARM	**35.9**	28.7	39.8	49.3	51.0	64.0	64.7	47.6
CMNeXt	1.9	39.0	0.1	45.9	4.9	55.8	59.6	29.6
GeminiFusion	9.3	38.0	3.9	45.7	50.5	69.8	71.8	41.3
MAGIC	29.3	44.7	41.4	51.2	51.6	65.8	66.0	50.0
StitchFusion	25.9	**47.6**	34.7	56.7	**57.5**	69.3	71.6	51.9
TokenFusion	10.1	39.6	4.2	49.6	18.0	49.9	55.1	32.4
A-MoM (ours)	24.2	45.0	**52.6**	**57.8**	56.2	**71.4**	**72.9**	**54.3**

Table 13: **Hyperparameters.** We report relevant hyperparameters for replicating our results.

Batch size	Epochs	Warmup epochs	Learning rate	Weight decay	AdamW β_1	AdamW β_2
4	500	50	$6 \cdot 10^{-5}$	0.01	0.9	0.999