EVIDENCESAM: HIERARCHICAL SOURCE EVIDENCE FUSION WITH SAM2 FOR MULTIMODAL SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal semantic segmentation benefits from complementary observations across heterogeneous sensors, yet the utility of each source can vary substantially across spatial regions and semantic categories. Existing approaches predominantly address this challenge through feature-level interaction, progressively compressing heterogeneous observations into a shared representation. However, source participation in representation learning does not necessarily reflect its contribution to the final semantic decision, particularly when adapting large segmentation foundation models. We introduce EvidenceSAM, a parameter-efficient SAM2-based framework that preserves source-specific evidence beyond feature fusion and explicitly models its role throughout semantic prediction. EvidenceSAM organizes multimodal reasoning into three complementary stages: Cross-Modal Feature Fusion (CMF) constructs a shared representation while retaining source-specific features; Risk-Decomposed Expert Aggregation (RDEA) evaluates source-associated semantic alternatives through asymmetric decision evidence; and Context-Guided Semantic Refinement (CGSR) incorporates class-level context into subsequent feature and prediction updates. This layered formulation separates representation-level source participation from prediction-level semantic contribution and contextual refinement, enabling heterogeneous sources to contribute differently throughout the prediction process. Built upon a largely frozen SAM2 backbone, EvidenceSAM requires only 12.9M trainable parameters for multimodal adaptation. Under four-modality settings, EvidenceSAM achieves 55.46% mIoU on MCubeS and 67.71% mIoU on DeLiVER, establishing state-of-the-art performance on both benchmarks. These results demonstrate the effectiveness of preserving and reasoning over source-specific semantic evidence for parameter-efficient multimodal foundation-model adaptation.

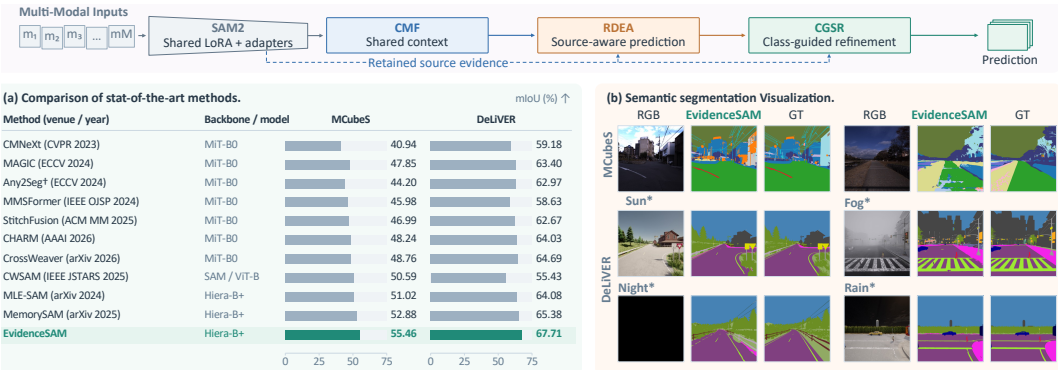

Figure 1: Overview and representative results of EvidenceSAM. EvidenceSAM preserves source-specific evidence beyond shared feature fusion and progressively performs cross-modal feature fusion, source-aware semantic aggregation, and context-guided refinement.

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.

1 INTRODUCTION

Multimodal semantic segmentation combines spatially aligned observations from heterogeneous sensors to recover dense semantic structure from complementary appearance, geometric, and material cues(Zhang et al., 2023b;a; Wu et al., 2022). Such information is particularly valuable when visual appearance alone is ambiguous, for example under illumination changes, adverse conditions, or material similarities(Zhang et al., 2021; Xu et al., 2021). However, the utility of each source is spatially and semantically dependent. A modality that is informative for one region or category may become redundant, noisy, or unreliable elsewhere(Peng et al., 2024; Zhao et al., 2025; Deregnaucourt et al., 2025). Effective multimodal segmentation therefore requires more than combining heterogeneous features. It must also determine how source-specific information participates in representation learning and semantic decisions.

Recent approaches have advanced multimodal representation learning through flexible cross-modal interaction, modality-agnostic learning, and multimodal pretraining (Zhang et al., 2023b; Zheng et al., 2024b;a; Yin et al., 2024; 2025; 2026). Attention-based, selective, and sparse fusion further improves information exchange across spatial locations and feature hierarchies (Jia et al., 2024; Li et al., 2025; Cai et al., 2025; Zheng et al., 2025; Liu et al., 2023). In parallel, parameter-efficient adaptation and cross-modal memory have extended segmentation foundation models to heterogeneous sensing inputs (Zhu et al., 2024; Liao et al., 2025b; Xiao et al., 2024; Curti et al., 2025; Xu et al., 2025). Despite these advances, multimodal adaptation remains largely centered on constructing a stronger shared representation. Once source information is absorbed into that representation, its role in the subsequent semantic decision is often no longer modeled explicitly. This raises a fundamental question: *should source participation in representation construction be treated as equivalent to its contribution to a semantic decision?*

We argue that feature participation and semantic contribution are different quantities. A source may contribute strongly to the shared representation without providing the most useful semantic alternative for the current prediction, while a source with lower feature-level participation may still provide useful prediction-level evidence. Our diagnostic analysis on MCubeS shows that feature-level source participation is only weakly associated with the prediction improvement provided by the corresponding source experts (Section 4.2). This motivates separating representation-level participation from prediction-level semantic contribution.

We introduce EvidenceSAM, a parameter-efficient SAM2-based framework for multimodal semantic segmentation that realizes this formulation, as illustrated in Figure 1. Rather than collapsing heterogeneous observations into a single fused representation, EvidenceSAM preserves source-specific evidence beyond feature fusion and reasons over it at multiple semantic levels. Cross-Modal Feature Fusion (CMF) constructs a shared multimodal representation while retaining individual source pyramids and adapts local source participation through Task-Aligned Source Redistribution (TASR). TASR uses bounded antisymmetric pairwise transfers and aligns the learned redistribution with downstream task responses. Risk-Decomposed Expert Aggregation (RDEA) maintains class-complete source predictions and evaluates their asymmetric decision evidence relative to the current semantic state rather than reducing source utility to a single feature-level reliability score. Context-Guided Semantic Refinement (CGSR) organizes source evidence into class-conditioned semantic states and propagates class context into subsequent feature and prediction updates. Together, the three components form a hierarchical source-evidence process that connects representation construction, source-aware semantic reasoning, and contextual refinement while preserving source identity across stages. EvidenceSAM adapts the pretrained SAM2 representation through shared LoRA and lightweight adapters while leaving most backbone parameters frozen.

EvidenceSAM achieves 55.46% mIoU on MCubeS and 67.71% mIoU on DeLiVER under four-modality settings. With only 12.9M trainable parameters for multimodal adaptation, EvidenceSAM consistently outperforms the compared Hiera-B+-based SAM2 adaptation methods on both benchmarks. The gains across material and driving-scene segmentation show that source-specific semantic evidence supports effective parameter-efficient multimodal adaptation.

Our contributions are summarized as follows. We introduce EvidenceSAM, a parameter-efficient SAM2 framework that formulates multimodal segmentation as hierarchical source-evidence fusion by separating feature-level source participation, prediction-level semantic contribution, and class-conditioned refinement while preserving source-specific evidence throughout the prediction process.

TASR provides task-aligned local source redistribution through bounded antisymmetric pairwise transfers, while RDEA and CGSR extend source reasoning from prediction-level aggregation to class-conditioned refinement. EvidenceSAM reaches 55.46% and 67.71% mIoU on MCubeS and DeLiVER, respectively, under four-modality settings, while requiring only 12.9M trainable parameters.

2 Related Work

Multimodal Semantic Segmentation. Multimodal semantic segmentation has evolved from pairwise RGB–X fusion toward architectures that support heterogeneous and dynamically available sensing streams (Wang et al., 2022; Zhang et al., 2023b; Yin et al., 2024). Recent methods learn modality-agnostic representations, flexible multimodal encoders, or modality-specific geometric priors to improve generalization across sensor combinations (Zheng et al., 2024b;a; Yin et al., 2025; 2026). Cross-modal attention, selective interaction, and sparse fusion further control information exchange across spatial locations and feature hierarchies (Jia et al., 2024; Li et al., 2025; Cai et al., 2025; Zhang et al., 2026; Wen et al., 2026). Recent work also studies modality imbalance and robustness to incomplete or degraded observations (Liao et al., 2025a; Zheng et al., 2025; Tan et al., 2026). These methods substantially advance multimodal representation learning, but their primary interaction occurs during representation construction. EvidenceSAM instead preserves source identity beyond shared feature fusion, allowing source-specific semantic alternatives to remain explicit during subsequent decision making. This separates representation-level source participation from prediction-level semantic contribution.

Multimodal Adaptation of SAM-family Models. SAM and SAM2 provide transferable segmentation representations that have motivated lightweight adaptation across semantic tasks and sensing modalities (Kirillov et al., 2023; Ravi et al., 2025; Xiong et al., 2024; Huang et al., 2024; Kweon & Yoon, 2024). For multimodal perception, existing approaches incorporate auxiliary observations through multimodal prompts or adapters (Xiao et al., 2024; Curti et al., 2025; Ren et al., 2025), modality-associated LoRA experts (Zhu et al., 2024), or cross-modal memory and semantic prototypes (Liao et al., 2025b). Recent work further extends SAM-family models to camera–LiDAR streams and other multimodal settings (Xu et al., 2025; Liu et al., 2026a; Pan et al., 2026). Parameter-efficient multimodal frameworks also adapt pretrained representations through modality-aware prompting and reusable cross-modal transfer (Cai et al., 2025; Liu et al., 2026b). EvidenceSAM follows this parameter-efficient paradigm through shared LoRA and lightweight adapters, but differs in where source-specific reasoning is retained. Rather than terminating multimodal interaction at backbone adaptation or shared feature fusion, EvidenceSAM preserves class-complete source predictions and reasons about their semantic contribution.

3 Method

3.1 Overall Framework

EvidenceSAM formulates multimodal semantic segmentation as *hierarchical source-evidence fusion*. Given M spatially aligned observations $\mathcal{X} = \{x_m\}_{m=1}^{M}$, SAM2-based pathways with shared LoRA and lightweight adapters extract multiscale source pyramids $\mathcal{F}_m = \{F_m^{\ell}\}_{\ell=1}^{L}$ on compatible grids. The model predicts C-class logits $z_{\text{out}} \in \mathbb{R}^{C\times H\times W}$.

As shown in Figure 2, EvidenceSAM separates multimodal reasoning according to the information available at different stages of prediction. Before semantic decoding, the model only has access to source features, availability, and quality estimates. Cross-Modal Feature Fusion (CMF) therefore operates at the representation level, determining how each source participates in the shared feature while retaining the individual source pyramids. After semantic decoding, the fused base prediction and source-associated semantic alternatives become available. Risk-Decomposed Expert Aggregation (RDEA) uses this new prediction-level information to evaluate whether a source-associated alternative should affect the current semantic decision. After aggregation, source evidence may still have different value across semantic classes and spatial regions. Context-Guided Semantic Refinement (CGSR) therefore organizes retained evidence by class and propagates class-conditioned feature and prediction updates.

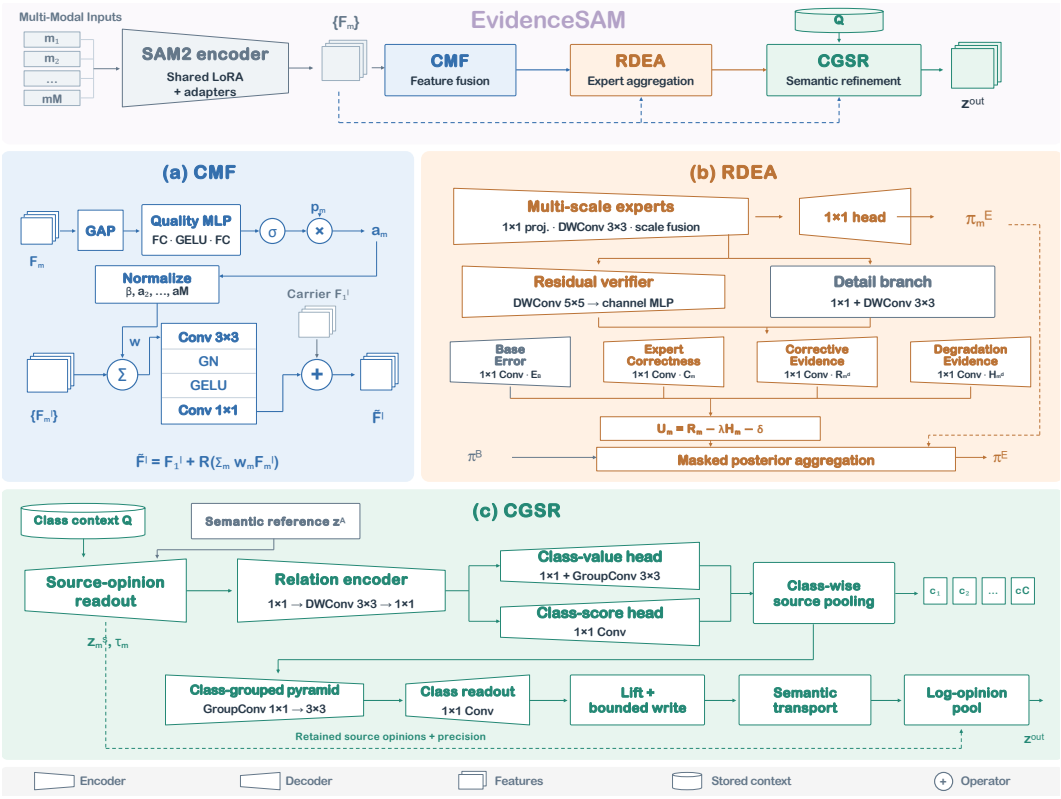

Figure 2: Architecture of EvidenceSAM. CMF builds the shared representation, RDEA evaluates prediction-level source contribution, and CGSR performs class-conditioned refinement.

The three stages address different roles. CMF determines how a source enters the shared representation. RDEA determines whether a source-associated semantic alternative should contribute to the current prediction. CGSR determines how the aggregated evidence should refine different semantic classes. This ordering follows the information available along the prediction path rather than assigning all three roles to a single source weight. Source identity is retained across stages so that feature fusion does not remove information needed for later semantic reasoning.

Stream m_1 serves as the residual carrier. For each auxiliary source, the support signal is defined as:

$$a_m = p_m q_m, \tag{1}$$

where p_m denotes source availability and $q_m \in [0, 1]$ denotes estimated source quality. CMF constructs the shared pyramid $\widetilde{\mathcal{F}}$ while retaining $\{\mathcal{F}_m\}_{m=1}^{M}$ for later semantic reasoning.

3.2 Cross-Modal Feature Fusion

CMF provides the shared representation used by the later semantic stages. A natural starting point is to weight sources by availability and estimated quality, reducing the influence of unavailable or low-quality observations. However, a single sample-level allocation cannot capture spatial changes in source usefulness. An auxiliary source may be informative near object or material boundaries but provide little additional evidence in a homogeneous region.

CMF therefore combines support-normalized fusion with *Task-Aligned Source Redistribution (TASR)*. Support-normalized fusion establishes a stable allocation from source availability and quality. TASR then makes bounded local changes to this allocation according to the downstream task response. CMF modifies source participation in the shared representation while retaining source pyramids for later semantic reasoning.

Support-normalized Fusion. At pyramid level ℓ, with the level index omitted for clarity, let $\beta = \max(b, 0)$ denote the nonnegative carrier weight. The source allocation and shared context are defined as:

$$\begin{aligned} d &= \max\left(\beta + \sum_{m=2}^{M} a_m, \epsilon\right), \\ w_1 &= \beta/d, \qquad w_m = a_m/d, \quad m \geq 2, \\ V_0 &= \sum_{m=1}^{M} w_m F_m, \\ \widetilde{F}_0 &= F_1 + \mathcal{R}(V_0), \end{aligned} \tag{2}$$

where $\mathcal{R}$ consists of a 3×3 convolution, group normalization, GELU, and a 1×1 projection. The allocation satisfies $w_m \geq 0$ and $\sum_m w_m \leq 1$. This construction gives the shared representation a stable reference before local redistribution. The residual carrier preserves the primary stream, while the normalized source terms control how much auxiliary information is introduced.

Task-aligned Source Redistribution. The initial allocation w_m captures source availability and sample-level quality but is shared across spatial locations. TASR introduces spatial adaptation at a coarse level $\ell_\star$.

We use pairwise redistribution instead of predicting a new independent weight for every source. Independent reweighting can change both relative allocation and total source support. Pairwise transfer instead couples an increase in one source with a decrease in another, keeping the update tied to the support-normalized allocation.

The source and context descriptors are:

$$h_m = P_\theta \mathcal{A}(F_m), \qquad c = P_\theta \mathcal{A}(V_0), \qquad \widehat{w}_m = \mathcal{A}(w_m), \tag{3}$$

where $\mathcal{A}$ denotes average pooling to the scoring grid and P_θ is a shared projection.

A shared scorer evaluates both orders of each source pair:

$$\begin{aligned} b_{mn} &= g_\theta([h_m, h_n, c, \widehat{w}_m, \widehat{w}_n]), \\ e_{mn} &= \mathcal{U}\left[\tanh\left(\frac{b_{mn} - b_{nm}}{2}\right)\right], \qquad e_{nm} = -e_{mn}, \end{aligned} \tag{4}$$

where $\mathcal{U}$ resizes the exchange field to the fusion grid. The antisymmetric form gives each source pair a transfer direction and bounds $|e_{mn}| \leq 1$.

For a redistribution scale $\rho \in [0, 1]$, the pairwise transfer and updated allocation are:

$$J_{mn} = \rho w_m w_n e_{mn}, \qquad w'_m = w_m + \sum_{n \neq m} J_{mn}. \tag{5}$$

The resulting update to the shared context is:

$$\begin{aligned} \Delta V_\theta &= \sum_{m<n} J_{mn}(F_m - F_n), \\ V_\theta &= V_0 + \Delta V_\theta, \\ \widetilde{F}_\theta^{\ell_\star} &= F_1^{\ell_\star} + \mathcal{R}_{\ell_\star}(V_\theta). \end{aligned} \tag{6}$$

Other pyramid levels follow Eq. equation 2. TASR changes source participation in the shared representation without altering the retained source pyramids. By antisymmetry, the redistribution preserves the total allocation mass, keeps source weights nonnegative, and bounds the resulting feature change. The corresponding bounds are given in the appendix. The scorer's final layer is zero-initialized, so $V_\theta = V_0$ initially.

3.3 Risk-Decomposed Expert Aggregation

CMF determines how source features participate in representation construction, but this does not determine whether a source-associated semantic prediction should influence the final decision. This

distinction becomes observable after semantic decoding, when the fused base prediction and source-associated alternatives are available, and is supported by the diagnostic analysis in Section 4.2.

RDEA addresses this prediction-level problem. Rather than reusing feature-fusion weights as semantic weights, RDEA retains class-complete source alternatives and evaluates each alternative relative to the fused base prediction. The aggregation decision therefore depends on the semantic state being modified rather than on source reliability alone.

Source-associated Semantic Alternatives. The fused representation produces the base prediction:

$$z_B = \mathcal{D}_B(\widetilde{\mathcal{F}}), \qquad \pi_B = \mathrm{softmax}(z_B). \tag{7}$$

For each source $m \in \mathcal{S}_E$, the source expert receives fused and source-specific features, task features, source support, and detached base logits. Its bounded residual prediction is:

$$\begin{aligned} d_m &= \kappa \tanh(D_m/\kappa), \\ z_m^E &= \mathrm{sg}(z_B) + a_m \left(d_m - \langle \pi_B, d_m \rangle_C \mathbf{1}_C\right), \\ \pi_m^E &= \mathrm{softmax}(z_m^E), \end{aligned} \tag{8}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $\kappa > 0$ bounds the residual magnitude. Each expert therefore produces a complete alternative over the same C classes as the base prediction.

Asymmetric Decision Evidence. A source alternative can affect the current prediction in two different ways. It may correct an error in the base prediction, but it may also replace a correct base decision with an incorrect alternative. These outcomes are asymmetric because a confident source can still degrade an already reliable base prediction. RDEA therefore does not represent semantic contribution with a single confidence score. It models corrective and degradation evidence separately and combines them when forming the decision utility.

Posterior entropy, confidence margins, disagreement, and projected visual features produce bounded factors E_B, C_m, R_m^d, and H_m^d. The decision evidence is defined as:

$$\begin{aligned} R_m &= \left[\max(E_B C_m R_m^d, \epsilon_r)\right]^{1/3}, \\ H_m &= \left[\max((1-E_B)(1-C_m) H_m^d, \epsilon_r)\right]^{1/3}, \\ U_m &= R_m - \lambda H_m - \delta. \end{aligned} \tag{9}$$

Here E_B and C_m describe the base and expert states, while R_m^d and H_m^d encode paired corrective and degradation evidence. The geometric form combines these factors on a common scale. The coefficient λ controls the degradation penalty and δ sets the selection margin. These factors are learned decision variables rather than calibrated probabilities.

A learned source preference t_m is combined with the utility:

$$s_m = t_m + \frac{U_m}{T_s} + \log \max(R_m, \epsilon_r). \tag{10}$$

A source is eligible when $a_m > 0$ and $U_m > 0$. After masking ineligible sources, the highest-scoring source is selected. The aggregated posterior is:

$$\begin{aligned} \chi_m &= \mathbf{1}[m = \arg\max_j \bar{s}_j]\, \mathbf{1}[a_m > 0, U_m > 0], \\ \alpha_m &= \eta \chi_m, \qquad \alpha_B = 1 - \sum_m \alpha_m, \\ \pi_E &= \alpha_B \pi_B + \sum_{m \in \mathcal{S}_E} \alpha_m \pi_m^E. \end{aligned} \tag{11}$$

If no source has positive decision utility, RDEA retains the base prediction rather than forcing auxiliary contribution. Training uses a masked soft assignment as a straight-through surrogate for the hard selection.

3.4 Context-Guided Semantic Refinement

RDEA determines whether source-associated semantic alternatives should contribute to the current prediction, but its aggregation does not by itself model how the retained evidence should be distributed

across semantic classes and feature levels. This matters in dense prediction because the same source can support different classes differently across regions.

CGSR addresses this remaining class-dependent structure. It organizes the semantic state produced by RDEA into class-conditioned source evidence and propagates this evidence to both the feature and prediction spaces. The stored class context Q provides the semantic reference for this process, while the source opinions retain information that would be lost if refinement operated only on the aggregated prediction.

Let z_A denote the semantic reference entering CGSR and Q the stored class context. With $\pi_A = \mathrm{softmax}(z_A)$, each source opinion is decoded as:

$$(z_m^S, \tau_m) = \mathcal{D}_m^S(x_m, \mathcal{F}_m, \pi_A; Q), \qquad m \in \mathcal{S}_S, \tag{12}$$

where z_m^S is a class-complete source opinion and τ_m is its precision field.

Class-factorized Evidence Encoding. A single source coefficient would assign the same semantic role to a source across all classes. CGSR instead factorizes source evidence by class. For each class, CGSR separates source value from contribution strength, allowing both source-specific and class-specific variation.

At level ℓ, a posterior-conditioned encoder derives class-specific source values $v_{m,c}^\ell$ and coefficients $\nu_{m,c}^\ell$ from source features, posterior relations, precision, and support. The class-wise source state is:

$$S_c^\ell = \sum_{m \in \mathcal{S}_S} \nu_{m,c}^\ell v_{m,c}^\ell. \tag{13}$$

A class-grouped top-down hierarchy then propagates the state across feature levels:

$$G_c^\ell = \mathcal{G}_c^\ell \left([S_c^\ell, \mathrm{Up}(G_c^{\ell+1})] \right). \tag{14}$$

The resulting class-space increments are written back to the shared features:

$$F^{\star,\ell} = \widetilde{F}^\ell + \mathcal{B}_\ell \left(L_\ell \boldsymbol{\delta}^\ell; \widetilde{F}^\ell \right), \tag{15}$$

where L_ℓ maps class-space increments to feature channels and $\mathcal{B}_\ell$ bounds the update magnitude.

Semantic Transport and Final Fusion. The refinement should change relative class evidence rather than add the same offset to every class logit. We therefore remove the class-common component of the transport signal before updating the semantic reference:

$$T = T_{\mathrm{raw}} - \frac{1}{C} \sum_c T_{\mathrm{raw},c} \mathbf{1}_C, \qquad z_J = z_A + \gamma T. \tag{16}$$

Here γ is an image-level source-support factor.

The refined reference and source opinions are combined as:

$$z_{\mathrm{out}} = \mathcal{P} \left(z_J, \{z_m^S, \tau_m\}_{m \in \mathcal{S}_S}; \tau_J \right), \tag{17}$$

where $\mathcal{P}$ denotes precision-weighted log-opinion pooling and τ_J is the precision of the refined reference. The definitions of γ and the pooling operation are provided in the appendix.

4 Experiments

4.1 Main Results

We compare EvidenceSAM with multimodal segmentation methods and SAM-family adaptation methods on MCubeS and DeLiVER. All results use four modalities. MCubeS uses RGB, AoLP, DoLP, and NIR, while DeLiVER uses RGB, depth, event, and LiDAR. For non-SAM methods, we report MiT-B0 configurations to keep the comparison consistent. SAM-family methods use the backbones listed in Table 1.

EvidenceSAM reaches 55.46% mIoU on MCubeS and 67.71% mIoU on DeLiVER. Among the compared methods using the same SAM2 / Hiera-B+ backbone, MemorySAM obtains 52.88% and 65.38% mIoU. EvidenceSAM improves these results by 2.58 and 2.33 percentage points, respectively. Compared with MLE-SAM, the gains are 4.44 points on MCubeS and 3.63 points on DeLiVER. Under the evaluated MiT-B0 and SAM-family comparison settings, EvidenceSAM achieves state-of-the-art performance on both benchmarks.

Table 1: Four-modality comparison on MCubeS and DeLiVER. Results are mIoU (%). Bold marks the best result under the reported comparison settings.

Method	Venue / Year	Backbone	MCubeS	DeLiVER
CMNeXt (Zhang et al., 2023b)	CVPR 2023	MiT-B0	40.94	59.18
MAGIC (Zheng et al., 2024b)	ECCV 2024	MiT-B0	47.85	63.40
Any2Seg (Zheng et al., 2024a)	ECCV 2024	MiT-B0	44.20	62.97
MMSFormer (Reza et al., 2024)	IEEE OJSP 2024	MiT-B0	45.98	58.63
StitchFusion (Li et al., 2025)	ACM MM 2025	MiT-B0	46.99	62.67
CHARM (Wen et al., 2026)	AAAI 2026	MiT-B0	48.24	64.03
CrossWeaver (Zhang et al., 2026)	arXiv 2026	MiT-B0	48.76	64.69
CWSAM (Pu et al., 2025)	IEEE JSTARS 2025	SAM / ViT-B	50.59	55.43
MLE-SAM (Zhu et al., 2024)	arXiv 2024	SAM2 / Hiera-B+	51.02	64.08
MemorySAM (Liao et al., 2025b)	arXiv 2025	SAM2 / Hiera-B+	52.88	65.38
EvidenceSAM (Ours)	–	SAM2 / Hiera-B+	**55.46**	**67.71**

4.2 Component and Diagnostic Analysis

We conduct a fixed-weight component analysis on MCubeS using the same 102 test images and identical inference settings. Each variant disables one component function without retraining. Removing CMF feature injection causes the largest drop, reducing mIoU from 55.46% to 51.94%. Bypassing RDEA reduces mIoU to 53.98%, while disabling CGSR gives 54.53%. The results show that representation construction, source-aware semantic aggregation, and class-conditioned refinement each contribute to the full model.

Table 2: Component analysis on MCubeS.

Variant	mIoU	Δ
w/o CMF injection	51.94	-3.52
w/o RDEA	53.98	-1.48
w/o CGSR	54.53	-0.93
Full	**55.46**	0.00

Feature Participation and Semantic Contribution. We test whether initial CMF weights averaged across pyramid levels reflect source-expert mIoU improvement over the fused base. Across 102 MCubeS test images, the Pearson correlations are -0.009, 0.038, and -0.022 for AoLP, DoLP, and NIR, respectively. All three 95% bootstrap confidence intervals include zero. Replacing RDEA's decision rule with CMF-weight or confidence-margin selection reduces mIoU by about 0.19 percentage points. Both replacements always admit an expert, so the comparisons reflect changes in both selection and admission. Together with the 1.48-point drop after bypassing RDEA, these results support separate modeling of feature participation and prediction-level semantic contribution in this setting (Appendix 11).

Internal Mechanism Analysis. Figure 3 visualizes the internal states of RDEA, TASR, and CGSR on the same MCubeS examples. Figure 3(a) shows that RDEA produces spatially distinct corrective and degradation evidence, leading to selective source contributions. Figure 3(b) shows how TASR maintains contributions from multiple sources while adapting their participation before semantic reasoning. Figure 3(c) shows class-conditioned evidence, localized feature updates, and semantic transport in CGSR. Together, these visualizations complement Table 2 by showing how the three components operate on different internal signals.

Qualitative Results

Figure 4 shows representative results on MCubeS and DeLiVER. EvidenceSAM preserves the main scene structure and semantic boundaries across the two sensor settings. The DeLiVER examples also show coherent predictions when event and LiDAR observations are sparse. Additional DeLiVER examples appear in Figure 5.

4.3 Model Complexity and Runtime

EvidenceSAM contains 119.28M parameters, of which 12.9M are trainable. Shared LoRA and lightweight adapters support multimodal adaptation while most SAM2 backbone parameters remain frozen. At an output resolution of 1024×1024 and batch size one, forward inference takes 534.85 ms per image on an RTX 5090 D and uses 5.35 GiB of peak allocated tensor memory. These measurements use the current unpruned implementation.

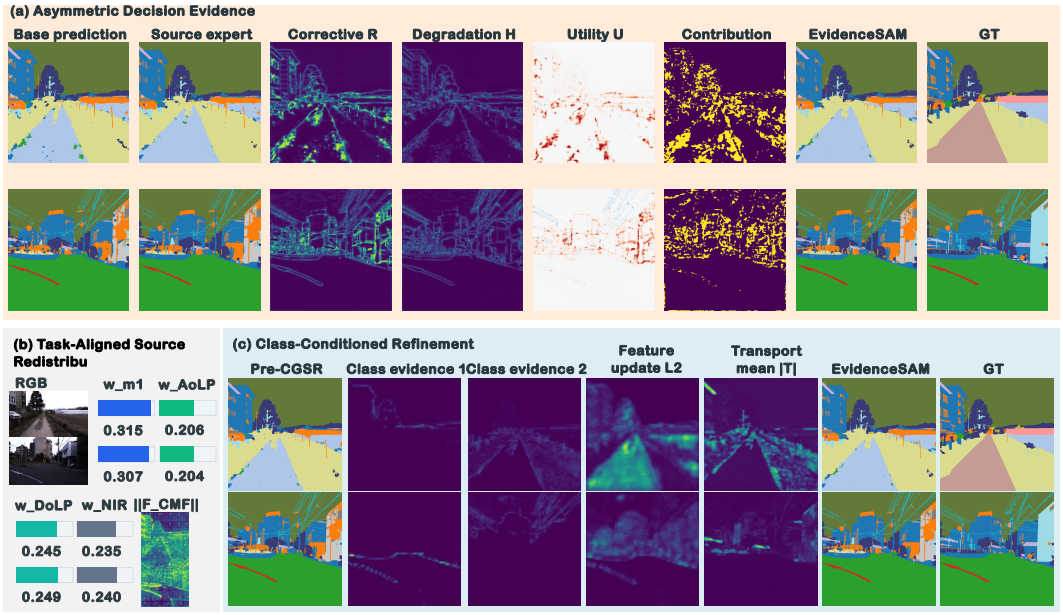

Figure 3: Internal mechanism visualization on MCubeS. (a) RDEA visualizes asymmetric decision evidence and source contribution. (b) TASR shows source allocation and the CMF response. (c) CGSR shows class-conditioned evidence, feature updates, and semantic transport.

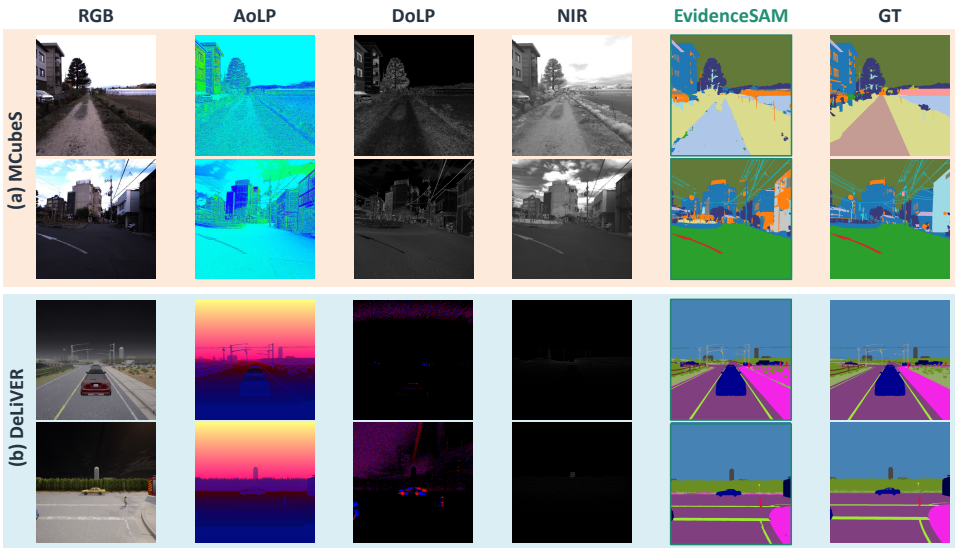

Figure 4: Qualitative results on MCubeS and DeLiVER.

5 Conclusion

We presented EvidenceSAM, a parameter-efficient SAM2 framework that preserves source-specific evidence beyond feature fusion and separates representation-level source participation, prediction-level semantic contribution, and class-conditioned refinement. EvidenceSAM achieves 55.46% mIoU on MCubeS and 67.71% mIoU on DeLiVER, reaching state-of-the-art performance under the evaluated comparison settings. These results support hierarchical source-evidence reasoning for parameter-efficient multimodal adaptation of segmentation foundation models.

Ethics statement

The work uses public or publicly described multimodal segmentation benchmarks and does not involve human-subject data collection. Multimodal perception can be relevant to robotics and autonomous systems, but the source-participation and expert-contribution signals in this paper are not calibrated safety certificates and should not be used as such without task-specific validation.

Reproducibility Statement

The main paper specifies the EvidenceSAM architecture and core inference path. The appendix provides the training objectives, implementation details, and benchmark configuration notes.

AI Use Statement

Generative AI tools were used solely for language editing and grammar checking to improve the clarity and readability of the manuscript.

References

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10587–10598. IEEE, 2025.

Iacopo Curti, Pierluigi Zama Ramirez, Alioscia Petrelli, and Luigi Di Stefano. Multimodal sam-adapter for semantic segmentation. *IEEE Access*, 2025.

Lucas Deregnaucourt, Hind Laghmara, Alexis Lechervy, and Samia Ainouz. A conflict-guided evidential multimodal fusion for semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1373–1382. IEEE, 2025.

Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. Segment and caption anything. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13405–13417. IEEE, 2024.

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 *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19499–19509. IEEE, 2024.

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*, pp. 1308–1317, 2025.

Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1567–1577. IEEE, 2025a.

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*, 2025b.

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 *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 8081–8090. IEEE, 2023.

Jiyuan Liu, Jia Lin, Xiaofei Zhou, Runmin Cong, Deyang Liu, and Zhi Liu. M^4-sam: Multi-modal mixture-of-experts with memory-augmented sam for rgb-d video salient object detection. *arXiv preprint arXiv:2605.11760*, 2026a.

Yajing Liu, Yumeng Zhang, Yue Si, Baojie Fan, and Jiandong Tian. Decoupled and reusable adaptation for efficient cross-modal transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 812–822, 2026b.

Jiancheng Pan, Runze Wang, Tianwen Qian, Mohammad Mahdi, Yanwei Fu, Xiangyang Xue, Xiaomeng Huang, Luc Van Gool, Danda Pani Paudel, and Yuqian Fu. V^ 2-sam: Marrying sam2 with multi-prompt experts for cross-view object correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16910–16919, 2026.

Fengguang Peng, Zihan Ding, Ziming Chen, Gang Wang, Tianrui Hui, Si Liu, and Hang Shi. Region-adaptive and context-complementary cross modulation for rgb-t semantic segmentation. *Pattern Recognition*, 147:110092, 2024.

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*, 18:4791–4804, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Guangyu Ren, Hengyan Liu, Michalis Lazarou, and Tania Stathaki. Multi-modal segment anything model for camouflaged scene segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19882–19892. IEEE, 2025.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Mmsformer: Multimodal transformer for material and semantic segmentation. *IEEE Open Journal of Signal Processing*, 5: 599–610, 2024.

Jiaqi Tan, Xu Zheng, and Yang Liu. Towards robust multi-modal semantic segmentation with teacher-student framework and hybrid prototype distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27665–27675, 2026.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12176–12185. IEEE, 2022.

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, pp. 10603–10611, 2026.

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

Aoran Xiao, Weihao Xuan, Heli Qi, Yun Xing, Naoto Yokoya, and Shijian Lu. Segment anything with multiple modalities. *arXiv preprint arXiv:2408.09085*, 2024.

Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16111–16121. IEEE, 2024.

Jiangtao Xu, Kaige Lu, and Han Wang. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognition Letters*, 146:179–184, 2021.

Jianyun Xu, Song Wang, Ziqian Ni, Chunyong Hu, Sheng Yang, Jianke Zhu, and Qiang Li. Sam4d: Segment anything in camera and lidar streams. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28535–28545. IEEE, 2025.

Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. Dformerv2: Geometry self-attention for rgbd semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19345–19355. IEEE, 2025.

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. *Advances in Neural Information Processing Systems*, 38:142674–142695, 2026.

Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *International Conference on Learning Representations*, volume 2024, pp. 51803–51825, 2024.

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*, 24(12):14679–14694, 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1136–1147. IEEE, 2023b.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2633–2642. IEEE, 2021.

Zelin Zhang, Kedi Li, Huiqi Liang, Chuanzhi Xu, Tao Zhang, and Weidong Cai. Crossweaver: Cross-modal weaving for arbitrary-modality semantic segmentation. *arXiv preprint arXiv:2604.02948*, 2026.

Shenlu Zhao, Ziniu Jin, Qiang Jiao, Qiang Zhang, and Jungong Han. Resolving semantic conflicts in rgb-t semantic segmentation. *Pattern Recognition*, 162:111398, 2025.

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

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*, pp. 192–212. Springer, 2024b.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21166–21176. IEEE, 2025.

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.

6 Cross-Modal Feature Fusion and Task-Aligned Source Redistribution

6.1 Notation and Support-Normalized Fusion

CMF builds a shared feature pyramid and retains the source pyramids for RDEA and CGSR. CMF combines support-normalized fusion and Task-Aligned Source Redistribution (TASR). Let F_m^ℓ denote the feature of source m at pyramid level ℓ, and let stream $m = 1$ be the residual carrier. For auxiliary sources, p_m denotes availability, $q_m \in [0, 1]$ denotes estimated quality, and $a_m = p_m q_m$ is the support. The quality estimator uses global average pooling followed by a two-layer MLP with GELU and a sigmoid output.

At a fixed level, omitting ℓ for clarity, define

$$\begin{aligned} \beta &= \max(b, 0), \qquad d = \max\left(\beta + \sum_{m=2}^{M} a_m, \epsilon\right), \\ w_1 &= \beta/d, \qquad w_m = a_m/d \quad (m \geq 2), \\ V_0 &= \sum_{m=1}^{M} w_m F_m, \qquad \widetilde{F}_0 = F_1 + \mathcal{R}(V_0). \end{aligned} \tag{18}$$

Here $\epsilon > 0$ is the denominator floor, and $\mathcal{R}$ contains a 3×3 convolution, group normalization, GELU, and a 1×1 projection. The initial weights are shared across spatial locations in each image. They are nonnegative and have total mass at most one.

6.2 Pairwise Scoring and Spatial Redistribution

At the coarse level $\ell^\star$, TASR adapts this initial allocation spatially. Average pooling $\mathcal{A}$ maps features to a scoring grid, and a shared projection P_θ produces descriptors

$$h_m = P_\theta \mathcal{A}(F_m), \qquad c = P_\theta \mathcal{A}(V_0), \qquad \widehat{w}_m = \mathcal{A}(w_m). \tag{19}$$

A shared pairwise scorer evaluates both orders of a pair:

$$\begin{aligned} b_{mn} &= g_\theta([h_m, h_n, c, \widehat{w}_m, \widehat{w}_n]), \\ e_{mn} &= \mathcal{U}\left[\tanh\left(\frac{b_{mn} - b_{nm}}{2}\right)\right], \quad e_{nm} = -e_{mn}. \end{aligned} \tag{20}$$

The operator $\mathcal{U}$ resizes the field to the fusion grid while preserving its range, so $|e_{mn}(u)| \leq 1$ at every spatial location u. The symbol b_{mn} denotes a pairwise score and is distinct from the carrier parameter b in Eq. equation 18.

For $\rho \in [0, 1]$, define

$$J_{mn}(u) = \rho w_m w_n e_{mn}(u), \qquad w'_m(u) = w_m + \sum_{n \neq m} J_{mn}(u). \tag{21}$$

Positive J_{mn} transfers allocation from source n to source m. The updated context and fused feature are

$$\begin{aligned} \Delta V_\theta(u) &= \sum_{m<n} J_{mn}(u)\big(F_m(u) - F_n(u)\big), \\ V_\theta &= V_0 + \Delta V_\theta = \sum_m w'_m F_m, \\ \widetilde{F}_\theta^{\ell^\star} &= F_1^{\ell^\star} + \mathcal{R}_{\ell^\star}(V_\theta). \end{aligned} \tag{22}$$

Other pyramid levels use Eq. equation 18. The retained source pyramids are unchanged. Zero initialization of the scorer's final layer gives $b_{mn} = 0$, $e_{mn} = 0$, and $V_\theta = V_0$ initially. Downstream task losses train the pairwise scorer without a separate TASR loss.

6.3 Mass Preservation, Non-negativity, and Bounded Updates

The following properties hold at each location on the fusion grid. Write $S = \sum_m w_m \in [0, 1]$, and assume $\rho \in [0, 1]$ and $|e_{mn}| \leq 1$. The proofs also cover the case where the denominator floor in Eq. equation 18 is active and $S < 1$.

Antisymmetry and mass preservation. Since $e_{nm} = -e_{mn}$ and $w_m w_n = w_n w_m$, we have $J_{nm} = -J_{mn}$. Consequently,

$$\sum_m w'_m = \sum_m w_m + \sum_{m<n} (J_{mn} + J_{nm}) = S. \tag{23}$$

TASR preserves the initial total weight, including when $S < 1$.

Non-negativity and support preservation. For each source,

$$\begin{aligned} |w'_m - w_m| &\leq \rho w_m \sum_{n \neq m} w_n = \rho w_m (S - w_m), \\ w'_m &\geq w_m \big[1 - \rho(S - w_m)\big] \geq 0. \end{aligned} \tag{24}$$

The last inequality follows from $S - w_m \leq 1$. Together with mass preservation, this yields $0 \leq w'_m \leq S \leq 1$. If $w_m = 0$, all transfers involving source m are zero and $w'_m = 0$. Thus a source with zero initial weight keeps zero weight in CMF.

Allocation-update bound. Summing the first inequality in Eq. equation 24 gives

$$\sum_m |w'_m - w_m| \leq \rho \left(S^2 - \sum_m w_m^2 \right) \leq \rho S^2 \left(1 - \frac{1}{M} \right). \tag{25}$$

The second step uses $\sum_m w_m^2 \geq S^2/M$.

Feature-update bound. For any norm on the feature channels, define the source diameter $D(u) = \max_{m<n} \|F_m(u) - F_n(u)\|$. Then

$$\begin{aligned} \|\Delta V_\theta(u)\| &\leq \rho \sum_{m<n} w_m w_n \|F_m(u) - F_n(u)\| \\ &\leq \frac{\rho D(u)}{2} \left(S^2 - \sum_m w_m^2 \right) \\ &\leq \frac{\rho D(u) S^2}{2} \left(1 - \frac{1}{M} \right). \end{aligned} \tag{26}$$

This bound follows directly from the triangle inequality and $2 \sum_{m<n} w_m w_n = S^2 - \sum_m w_m^2$. In particular, identical source features produce no context change, even when their allocations change. The context update is also zero when $\rho = 0$, when $S = 0$, or when only one source has nonzero allocation. If $\mathcal{R}_{\ell^\star}$ is $L_{\mathcal{R}}$-Lipschitz on the feature states considered, the corresponding full-tensor bound is

$$\|\widetilde{F}_\theta^{\ell^\star} - \widetilde{F}_0^{\ell^\star}\| \leq L_{\mathcal{R}} \|\Delta V_\theta\|. \tag{27}$$

Only the bound on the fused feature requires this Lipschitz assumption. The context bound in Eq. equation 26 holds without it.

7 Risk-Decomposed Expert Aggregation

7.1 Source-Associated Semantic Alternatives

The shared pyramid produces a fused base prediction $z^B = D_B(\widetilde{F})$ with posterior $\pi^B = \mathrm{softmax}(z^B)$. For each $m \in \mathcal{S}_E$, the source expert receives fused and source-specific features,

task features, source support, and detached base logits. With D_m denoting its residual output, it forms

$$\begin{aligned} d_m &= \kappa \tanh(D_m/\kappa), \\ z_m^E &= \mathrm{sg}(z^B) + a_m \left(d_m - \langle \pi^B, d_m \rangle_C \mathbf{1}_C\right), \\ \pi_m^E &= \mathrm{softmax}(z_m^E). \end{aligned} \tag{28}$$

Here $\kappa > 0$, sg denotes stop-gradient, and $\langle \pi^B, d_m \rangle_C = \sum_c \pi_c^B d_{m,c}$. Each alternative covers all C classes. Subtracting the posterior-weighted class-common term centers the residual without changing its softmax relative to the uncentered residual. Since $|d_{m,c}| \le \kappa$, the centered residual has magnitude at most 2κ per class.

7.2 Corrective Evidence, Degradation Evidence, and Utility

Posterior entropy, confidence margins, disagreement, and projected visual features provide the bounded factors E_B, C_m, R_m^d, and H_m^d. The evidence and utility follow the main paper:

$$\begin{aligned} R_m &= \left[\max(E_B C_m R_m^d, \epsilon_r)\right]^{1/3}, \\ H_m &= \left[\max((1-E_B)(1-C_m) H_m^d, \epsilon_r)\right]^{1/3}, \\ U_m &= R_m - \lambda H_m - \delta, \\ s_m &= t_m + U_m/T_s + \log \max(R_m, \epsilon_r). \end{aligned} \tag{29}$$

Here t_m is a learned source preference, $T_s > 0$ is the utility scale, λ controls the degradation penalty, and δ is the selection margin. The factors E_B, C_m, R_m^d, and H_m^d are learned decision variables rather than calibrated probabilities.

7.3 Eligibility and Masked Posterior Aggregation

At each location, define $k_m = \mathbf{1}[a_m > 0,\ U_m > 0]$ and set $\overline{s}_m = s_m$ for eligible sources and $-\infty$ otherwise. If at least one source is eligible, the highest-scoring eligible source is selected. If no source is eligible, all source indicators are zero. Thus

$$\begin{aligned} \chi_m &= k_m \mathbf{1}[m = \arg\max_j \overline{s}_j], \\ \alpha_m &= \eta \chi_m, \qquad \alpha_B = 1 - \sum_{m \in \mathcal{S}_E} \alpha_m, \\ \pi^E &= \alpha_B \pi^B + \sum_{m \in \mathcal{S}_E} \alpha_m \pi_m^E. \end{aligned} \tag{30}$$

The no-eligible-source case is defined separately, so no argmax over an empty eligible set is needed. With mixing strength $\eta \in [0, 1]$, Eq. equation 30 is a convex posterior mixture and returns π^B when every source is ineligible. During training, a masked soft assignment supplies a straight-through surrogate for the hard selection. Writing this assignment as $\widetilde{\chi}$, the surrogate has the form $\mathrm{sg}(\chi - \widetilde{\chi}) + \widetilde{\chi}$. This expression uses hard selection in the forward pass and gradients from the soft assignment in the backward pass. Ineligible entries are masked, and the all-ineligible case has zero source assignment.

8 Context-Guided Semantic Refinement

8.1 Class-Factorized Evidence and Class-Grouped Hierarchy

Let z^A denote the semantic reference entering CGSR and $\pi^A = \mathrm{softmax}(z^A)$. The reference z^A is multimodal. For $m \in \mathcal{S}_S$, a source-opinion decoder produces

$$(z_m^S, \tau_m) = D_m^S(x_m, F_m, \pi^A; Q), \tag{31}$$

where Q is stored class context, z_m^S is a class-complete source opinion, and τ_m is its precision field. The expert index set $\mathcal{S}_E$ in RDEA and the opinion index set $\mathcal{S}_S$ are kept separate in the notation.

At level ℓ, the posterior-conditioned encoder uses source features, posterior relations, precision, and support to produce source values $v^{\ell}_{m,c}$ and contribution coefficients $\nu^{\ell}_{m,c}$ for each class. The encoder forms a state for each class, which a top-down hierarchy passes across feature levels:

$$S^{\ell}_{c} = \sum_{m \in \mathcal{S}_S} \nu^{\ell}_{m,c} v^{\ell}_{m,c}, \qquad G^{\ell}_{c} = \mathcal{G}^{\ell}_{c}\left([S^{\ell}_{c}, \mathrm{Up}(G^{\ell+1}_{c})]\right). \tag{32}$$

The class-factorized representation uses 16 channels per class. The hierarchy preserves class grouping while passing context across feature levels. The hierarchy produces class-space increments δ^{ℓ} that update the shared features:

$$F^{\star,\ell} = \widetilde{F}^{\ell} + \mathcal{B}_{\ell}\left(L_{\ell}\delta^{\ell}; \widetilde{F}^{\ell}\right), \tag{33}$$

where L_{ℓ} maps class-space increments to feature channels and $\mathcal{B}_{\ell}$ bounds the update magnitude. Source values and contribution strengths can therefore differ across both sources and classes.

8.2 Stored Class Context

For source m and class c, the context readout combines a stored vector $\mu_{m,c}$, its valid-pixel count $n_{m,c}$, and a learned query $q^{\mathrm{learn}}_{m,c}$:

$$\begin{aligned} \eta^{\mathrm{ctx}}_{m,c} &= \frac{n_{m,c}}{n_{m,c} + 4096}, \\ Q_{m,c} &= \mathrm{normalize}\left(\eta^{\mathrm{ctx}}_{m,c}\mu_{m,c} + (1 - \eta^{\mathrm{ctx}}_{m,c}) q^{\mathrm{learn}}_{m,c}\right), \\ s^{\mathrm{ctx}}_{m,c}(u) &= \frac{\langle \mathrm{normalize}(f_m(u)), Q_{m,c}\rangle}{0.2}. \end{aligned} \tag{34}$$

The superscript ctx distinguishes the context interpolation coefficient and similarity from RDEA's mixing strength and selection score. The source-opinion width is 96 and the context scale is 0.25. Class statistics use training labels resized by nearest-neighbor interpolation and include only pixels with positive source presence. Distributed training combines sums and counts across workers. The stored vectors and counts remain fixed during evaluation.

8.3 Source Support and Semantic Transport

The image-level source-support factor is

$$\gamma = 1 - \prod_{m \in \mathcal{S}_S} \left[1 - \mathrm{clip}\left(\langle p_m\rangle \langle q_m\rangle, 0, 1\right)\right]. \tag{35}$$

The means $\langle p_m\rangle$ and $\langle q_m\rangle$ are computed separately over the first channel of each support map. The source-support factor therefore uses the product of these means. The product over m includes available source pyramids. Missing support metadata uses a value of one, and an empty product equals one. Thus $\gamma \in [0, 1]$ and $\gamma = 0$ when every source-support factor is zero.

CGSR removes the class-common component of the transport signal:

$$T = T_{\mathrm{raw}} - \frac{1}{C}\left(\sum_{c} T_{\mathrm{raw},c}\right)\mathbf{1}_C, \qquad z^{J} = z^{A} + \gamma T. \tag{36}$$

Thus $\sum_c T_c(u) = 0$ at each location. This centering removes a degree of freedom that does not affect softmax probabilities and expresses the transport in terms of relative class evidence.

8.4 Precision-Weighted Log-Opinion Pooling

Let $\mathcal{U}$ be bilinear interpolation to the reference output grid, with `align_corners=False`. Define

$$\ell_{J,c} = \mathrm{logsoftmax}(z^{J})_c, \qquad \ell_{m,c} = \mathrm{logsoftmax}(\mathcal{U} z^{S}_{m})_c, \qquad \widehat{\tau}_{m,c} = \mathcal{U}\tau_{m,c}. \tag{37}$$

Source logits are interpolated before log-softmax. The reference precision is $\tau_J = 1$. To distinguish the precision-head output from source support a_m, denote that output by $r^{\mathrm{prec}}_{m,c}$. Source precision is

$$\begin{aligned} \overline{q}_m &= 0.5 + 0.5\,\mathrm{clip}(q_m, 0, 1), \\ \tau_{m,c} &= \min\{\mathrm{softplus}(r^{\mathrm{prec}}_{m,c}), 2\}\,\mathcal{U}_m(p_m)\,\mathcal{U}_m(\overline{q}_m), \end{aligned} \tag{38}$$

where $\mathcal{U}_m$ resizes support to the precision-head grid. Zero presence gives zero precision. The quality floor means that zero raw quality alone does not force opinion precision to zero. The factor γ in Eq. equation 35 instead uses raw quality. The source-opinion decoder uses source dropout probability 0.1 when it is in training mode.

The final pooling operation is

$$\begin{aligned} v_c &= \frac{\tau_J \ell_{J,c} + \sum_{m \in \mathcal{S}_S} \widehat{\tau}_{m,c} \ell_{m,c}}{\max\left(\tau_J + \sum_{m \in \mathcal{S}_S} \widehat{\tau}_{m,c}, 10^{-6}\right)}, \\ z_{\mathrm{out},c} &= v_c - \log \sum_{c'=1}^{C} \exp(v_{c'}). \end{aligned} \tag{39}$$

The denominator depends on class and location. This is a weighted combination of log opinions followed by normalization over classes, rather than an arithmetic average of logits or posterior probabilities. Pooling is computed in float32. If all source precisions are zero, the result is $\mathrm{logsoftmax}(z^J)$, preserving the refined reference posterior and its argmax. When $\gamma = 0$, semantic transport leaves z^A unchanged. Source opinions can still contribute through the precision weights because Eq. equation 38 applies a floor to quality.

9 Training Objectives and Implementation Details

9.1 Semantic Supervision

For logits z and target labels y, the semantic loss is

$$\mathcal{L}_{\mathrm{sem}}(z, y) = \mathcal{L}_{\mathrm{CE}} + 0.5\mathcal{L}_{\mathrm{Dice}} + 0.25\mathcal{L}_{\mathrm{OHEM}} + 0.2\mathcal{L}_{\partial}. \tag{40}$$

For the MCubeS configuration, all terms exclude ignore label 255 and label smoothing is zero. Cross-entropy is averaged over valid pixels. OHEM averages the largest $\max(1, \lfloor 0.25 N_{\mathrm{valid}} \rfloor)$ valid-pixel cross-entropies. The boundary term averages cross-entropy over pixels on either side of horizontal or vertical target-label changes, provided both pixels are valid. The boundary loss is zero if no such boundary exists. Dice loss averages over target classes present in the batch:

$$\mathcal{L}_{\mathrm{Dice}} = \frac{1}{|\mathcal{C}_y|} \sum_{c \in \mathcal{C}_y} \left(1 - \frac{2\sum_u \pi_c(u)\mathbf{1}[y(u) = c] + 1}{\sum_u \pi_c(u) + \sum_u \mathbf{1}[y(u) = c] + 1}\right), \tag{41}$$

where $\pi = \mathrm{softmax}(z)$ and the sums include valid pixels only.

Semantic supervision applies to the final output and the prediction branches active in a training stage. The stage loss is

$$\begin{aligned} \mathcal{L}^{(s)} = \mathcal{L}_{\mathrm{sem}}(z_{\mathrm{out}}, y) + \lambda_B^{(s)} \mathcal{L}_{\mathrm{sem}}(z^B, y) + \lambda_C^{(s)} \mathcal{L}_{\mathrm{sem}}(z^{\mathrm{cand}}, y) \\ + \frac{\lambda_E^{(s)}}{|\mathcal{S}_E|} \sum_{m \in \mathcal{S}_E} \mathcal{L}_{\mathrm{sem}}(z_m^E, y) + \lambda_R^{(s)} \mathcal{L}_{\mathrm{RDEA}}. \end{aligned} \tag{42}$$

Here z^{cand} denotes the supervised candidate output when that branch is active. The expert term averages complete semantic losses. In the MCubeS joint-adaptation configuration, the effective weights are

$$\lambda_B^{(s)} = 1, \quad \lambda_C^{(s)} = 0.5 s_t, \quad \lambda_E^{(s)} = 1 - 0.5 s_t, \quad \lambda_R^{(s)} = 0.5\mathbf{1}[s_t > 0], \tag{43}$$

with curriculum factor

$$s_t = \begin{cases} 0, & t < 1500, \\ \min\{1, (t - 1500 + 1)/1125\}, & t \geq 1500. \end{cases} \tag{44}$$

The variable t is the persistent curriculum counter before the forward pass. A resumed stage continues from its loaded counter. At $s_t = 1$, the candidate, expert, and RDEA coefficients are all 0.5. These coefficients apply to the joint-adaptation stage.

Table 3: MCubeS joint-adaptation configuration. The values describe this training stage.

Setting	Value
Configured epochs	100
Peak learning rate	5×10^{-5}
Per-device batch size	2
Optimizer	AdamW
Polynomial decay power	0.9
Warmup fraction / initial ratio	0.02 / 0.1
Weight decay	0.01
Bias and LayerNorm weight decay	0
Mixed-precision dtype	bfloat16
Gradient clipping	None
Run seed	123

9.2 RDEA Auxiliary Supervision

Let $\widehat{y}^B = \arg\max_c \pi_c^B$ and $\widehat{y}_m^E = \arg\max_c \pi_{m,c}^E$. The training targets are computed from detached predictions and ground-truth labels:

$$\begin{aligned} b_{\text{err}} &= \mathbf{1}[\widehat{y}^B \neq y], \qquad e_{m,\text{correct}} = \mathbf{1}[\widehat{y}_m^E = y], \\ r_m^{\text{tgt}} &= b_{\text{err}} e_{m,\text{correct}}, \qquad h_m^{\text{tgt}} = (1 - b_{\text{err}})(1 - e_{m,\text{correct}}). \end{aligned} \tag{45}$$

Repair occurs when the base prediction is incorrect and the expert is correct. Harm occurs when the base prediction is correct and the expert is incorrect. The joint-adaptation auxiliary objective is

$$\begin{aligned} \mathcal{L}_{\text{RDEA}} = 0.25\mathcal{L}_{B\text{-err}} + 0.25\mathcal{L}_{E\text{-correct}} + \mathcal{L}_{\text{repair}} + \mathcal{L}_{\text{harm}} \\ + 0.5\mathcal{L}_{\text{utility}} + 0.25\mathcal{L}_{\text{select}}. \end{aligned} \tag{46}$$

The first four terms are binary cross-entropies. Positive base-error, repair, and harm examples receive importance factors 2, 4, and 4, respectively. Expert correctness has no extra positive factor. Reductions use normalized inverse class-frequency weights over valid, available examples.

Utility supervision uses SmoothL1 with $\beta = 0.1$ and target $r_m^{\text{tgt}} - 1.25h_m^{\text{tgt}}$. Neutral pixels receive importance 0.1 and repair/harm pixels importance 1. The penalty 1.25 belongs to the utility training target. It does not specify the inference coefficient λ in Eq. equation 29. Source-selection cross-entropy is evaluated only where an available expert repairs the base. Among repairing experts, its target maximizes

$$g_m^{\text{tgt}} = \text{clip}\left(\tanh(\ell_B - \ell_m) - 0.02, -2, 2\right), \tag{47}$$

where ℓ_B and ℓ_m are detached per-pixel cross-entropies. Ties use the first maximizing source index. These targets use training labels. Evidence estimation at inference does not use labels.

9.3 Backbone and MCubeS Joint Adaptation

The model uses SAM2/Hiera-B+ with shared LoRA and lightweight adapters, leaving most pretrained backbone parameters frozen. The MCubeS implementation uses SAM2.1/Hiera-B+. Q/V LoRA has rank 4, scaling parameter 4, and dropout 0, without a restricted layer list. Table 3 gives the joint-adaptation settings.

10 Benchmark and Evaluation Details

Both benchmarks use four modalities. MCubeS uses RGB, angle of linear polarization (AoLP), degree of linear polarization (DoLP), and near-infrared (NIR). DeLiVER uses RGB, depth, event, and LiDAR. EvidenceSAM achieves 55.46% mIoU on MCubeS and 67.71% on DeLiVER.

The MCubeS evaluation has 20 classes and ignore label 255, applies a 192-column crop from the left side, and uses 1024×1024 prediction grids. Inference is in float32 without test-time augmentation. The semantic output is `mmsam2_semantic_logits`. The component and source-participation analyses use the same 102 test images, model weights, preprocessing, and persistent state. Dataset

Table 4: Initial CMF participation versus per-image expert mIoU change. Each source has 102 paired observations. Confidence intervals use 2000 image-bootstrap draws. The p-value uses a two-sided image-block permutation test with 2000 permutations and a plus-one correction.

Source	Pearson r	95% CI for r	p	Spearman correlation
AoLP	-0.0094	$[-0.1171,\ 0.1043]$	0.9465	0.0726
DoLP	0.0384	$[-0.1212,\ 0.1844]$	0.6827	-0.0062
NIR	-0.0215	$[-0.2196,\ 0.1636]$	0.8326	-0.1059

Table 5: Fixed-weight source-decision interventions on MCubeS. Changes are relative to the full model in percentage points. No retraining is used.

Decision rule	mIoU (%)	Δ mIoU
CMF-weight selection	55.2737	-0.1903
Confidence-margin selection	55.2726	-0.1914
Full EvidenceSAM	**55.4640**	**0.0000**

mIoU is computed after aggregating confusion matrices across images. Per-image diagnostic mIoU averages classes with nonzero union within that image.

The main comparison reports non-SAM methods in their MiT-B0 configurations and SAM-family methods with the backbones listed in the main table. The optimization and preprocessing settings in this appendix apply to MCubeS only.

11 Diagnostic Analysis of Source Participation

11.1 Feature Participation and Semantic Improvement

The diagnostic compares the initial, support-normalized CMF allocation with the mIoU improvement of each source expert over the fused base prediction. For test image i, the participation statistic is

$$\overline{w}_{i,m} = \frac{1}{|\mathcal{L}_i|} \sum_{\ell \in \mathcal{L}_i} w_{i,m}^{\ell}, \qquad \Delta_{i,m} = \mathrm{mIoU}_i(\widehat{y}_m^E) - \mathrm{mIoU}_i(\widehat{y}^B), \tag{48}$$

where $\mathcal{L}_i$ is the set of executed pyramid levels. The statistic uses w_m from Eq. equation 18, not the spatially redistributed w'_m or the exchange field e_{mn}. The expert and base predictions are scored against ground truth after inference. The diagnostic measures the association between initial CMF weight and expert improvement over the base. It does not measure the effect of removing an input source.

Table 4 reports the correlation results. All Pearson estimates are close to zero, and every interval includes zero. These observations do not establish statistical independence or exclude nonlinear associations. The source with the largest mean CMF weight also has the largest per-image expert improvement on 32 of 102 images (31.37%).

11.2 Fixed-Weight Decision and Component Interventions

The interventions in Table 5 replace the source-decision rule and retain the trained prediction paths and downstream computation. CMF-weight selection uses $\overline{w}_{i,m}$ from Eq. equation 48. Confidence-margin selection uses the confidence margins of source predictions.

Both replacement rules reduce mIoU by approximately 0.19 points. Both replacement rules always admit an expert in this evaluation with all four input modalities. RDEA retains the base prediction on 53.79% of valid pixels. The performance differences therefore reflect both source selection and whether an expert is admitted. They do not isolate the effect of separating corrective and degradation evidence.

Table 6 reports the main paper's component interventions under the same fixed-weight protocol.

Table 6: MCubeS component interventions without retraining. Changes are in percentage points relative to Full.

Variant	mIoU (%)	Δ mIoU
Without CMF feature injection	51.94	−3.52
Without RDEA	53.98	−1.48
Without CGSR	54.53	−0.93
Full EvidenceSAM	55.46	0.00

12 Model Complexity and Runtime

EvidenceSAM contains 119.28M total parameters, of which 12.9M are trainable. At 1024×1024 output resolution and batch size one, forward inference takes 534.85 ms per image on an RTX 5090 D and uses 5.35 GiB of peak allocated tensor memory. These measurements describe the current unpruned implementation.

13 Additional DeLiVER Qualitative Results

Figure 5 shows eight DeLiVER examples, with two examples each for Sun, Fog, Night, and Rain. The four input modalities provide different views of the same scene. RGB appearance varies with lighting and weather, depth shows scene geometry, and event and LiDAR observations provide sparse spatial cues. The predictions capture the main road layout and large scene regions across these examples, including scenes with low RGB visibility. Differences from the ground truth remain around some object boundaries and small objects. These examples illustrate the model's predictions under the four conditions without providing a separate quantitative comparison for each condition.

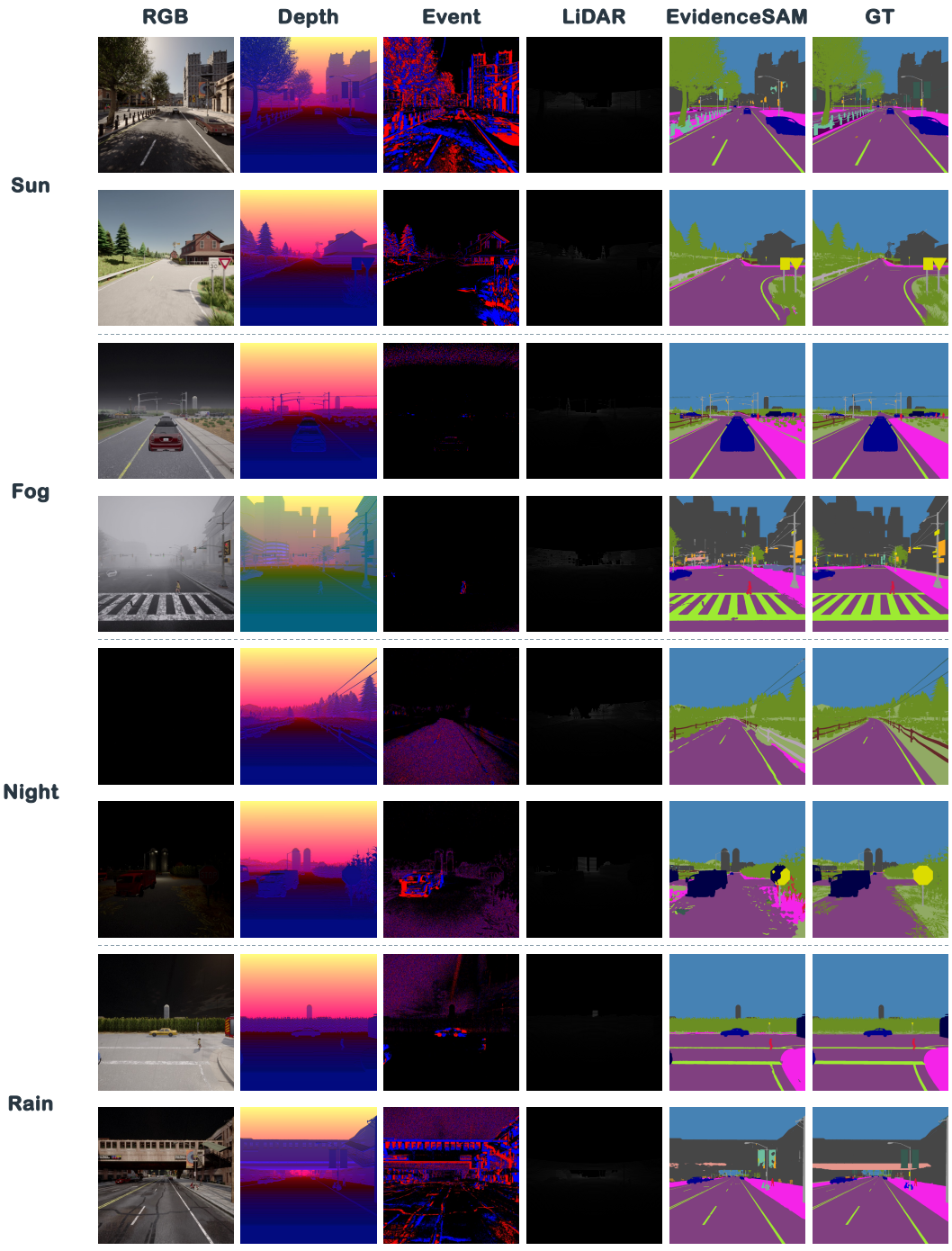

Figure 5: Additional qualitative results on DeLiVER. Each row shows one scene. Columns from left to right are RGB, depth, event, LiDAR, EvidenceSAM prediction, and ground truth (GT). Row groups correspond to Sun, Fog, Night, and Rain, with two examples per group.