GENFUSE: CROSS-SAMPLE GENETIC FUSION FOR INCOMPLETE MULTIMODAL TRAINING IN REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal remote sensing classifiers typically assume complete training pairs and handle missing modalities only at inference, limiting their applicability to incomplete multimodal training (IMT), where modality absence also occurs during training. We propose GenFuse, a cross-sample genetic fusion framework that generates a missing-modality proxy by recombining two complementary parents rather than relying solely on intra-sample mapping. The current sample's available feature serves as the semantic parent, while features retrieved from the training bank corresponding to the missing modality are aggregated into the modality parent. Feature-wise crossover and range-scaled mutation with a learnable offset produce the proxy, which is fused with the available feature for classification. The complete subset supervises proxy alignment, while genetic consistency anchors proxies for incomplete samples to their available features without ground-truth features from the missing modality. Experiments on three datasets under specific and probabilistic modality-missing protocols demonstrate superior classification performance.

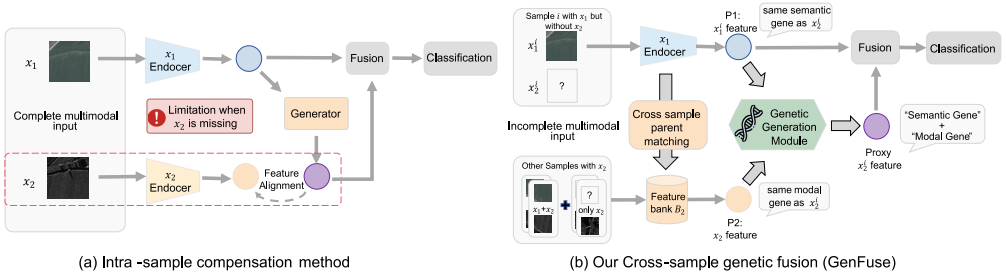
1 INTRODUCTION

Multimodal remote sensing integrates complementary observations from optical imagery, SAR, and LiDAR to improve representation learning beyond a single modality Guo & Gu (2025); Du et al. (2025); Qu et al. (2024b); Guo et al. (2024a); Liu et al. (2024). Fusion networks are commonly trained with complete, well-aligned modality pairs Zhao et al. (2024); Wang et al. (2024); Zhao et al. (2023), but sensor availability and acquisition conditions may leave one or more modalities unavailable at deployment Ke et al. (2025). This problem has motivated generative feature compensation Reza et al. (2025); Do et al. (2025); Qu et al. (2024a), prompt-based feature compensation Chen et al. (2026); Lee et al. (2023); Zhang et al. (2025a), and knowledge distillation Wei et al. (2023a); Sikdar et al. (2025); Xin et al. (2024). As illustrated in Figure 1a, most of these methods address missing modalities only during inference and assume complete modality pairs during training to provide same-sample supervision for compensation Qu et al. (2024a); Liu et al. (2023); Woo et al. (2023).

However, RS training collections may not contain all modalities for every sample. Differences in sensor availability, coverage, and acquisition conditions can leave some training samples incomplete while others retain complete modality pairs Xu et al. (2026); Han et al. (2025). This gives rise to incomplete multimodal training (IMT), where modality absence also occurs during training rather than only at inference. For an incomplete sample, the feature of the missing modality is unavailable, so generation, reconstruction, or proxy alignment cannot be supervised using the corresponding feature from that sample. Relying exclusively on complete samples leaves the available modality in each incomplete sample unused, reducing the data available for learning missing-modality compensation. The central challenge in IMT is to incorporate incomplete samples into training without access to their missing-modality features.

In contrast, Figure 1b illustrates the cross-sample strategy of GenFuse, which draws on two complementary sources that remain available for an incomplete sample. The available modality of the current sample yields a semantic parent that carries current-scene semantics but does not represent

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.



(a) Intra -sample compensation method (b) Our Cross-sample genetic fusion (GenFuse)

Figure 1: Comparison of (a) conventional intra-sample compensation using complete pairs and (b) GenFuse cross-sample proxy generation.

the missing modality. The feature bank corresponding to the missing modality provides representations of that modality from other training samples, which form a modality parent but may not fully match the current scene. Inspired by the recombination of parental information in biological inheritance Yelmen & Jay (2023); Huang et al. (2024b), GenFuse recombines these parents to generate a proxy for the missing modality, which is subsequently used for classification.

To realize this recombination, GenFuse comprises Cross-Sample Parental Matching (CSPM) and the Genetic Generation Module (GGM). CSPM retrieves multiple candidates from the feature bank corresponding to the missing modality and constructs the modality parent through attention-weighted aggregation. Rather than using this parent directly as the proxy, GGM recombines it with the semantic parent through feature-wise crossover and then applies range-scaled mutation using a learnable offset scaled by the feature range of the corresponding bank. Unlike conventional gated fusion between two observed modalities, this process combines the semantics of the current sample with representations of the missing modality drawn from other training samples. The resulting proxy is fused with the available feature to classify incomplete samples. During training, complete samples provide supervision for bidirectional proxy alignment, while the genetic consistency loss encourages semantic consistency between the proxy and the available feature of an incomplete sample without requiring its missing-modality feature.

The main contributions of this work are summarized as follows:

- We propose GenFuse for remote sensing IMT, which constructs two complementary parents from current-sample semantics and cross-sample features of the missing modality.
- We develop CSPM to construct the modality parent and GGM to generate the proxy through feature-wise crossover and range-scaled mutation, together with training objectives tailored to complete and incomplete samples.
- Extensive experiments on three remote sensing datasets under specific and probabilistic modality-missing protocols demonstrate the effectiveness of GenFuse, with further analyses of retrieval validity, alternative proxy strategies, proxy-feature alignment, and computational efficiency.

2 Related Work

2.1 Multimodal Learning with Missing Modalities

Multimodal learning integrates complementary signals from heterogeneous sources Pipoli et al. (2025); Swetha et al. (2023). Many missing-modality methods assume complete training pairs and address modality absence at inference Wei et al. (2023c); Ma et al. (2022); Sun et al. (2026). LDS2AE reconstructs masked features Qu et al. (2024a), while language-guided prompting compensates for unavailable inputs Huang et al. (2024a). DPMamba transfers knowledge from a full-modality teacher Yang et al. (2025), whereas dynamic selection adapts the active modalities at inference Du et al. (2026). Their reliance on complete pairs for reconstruction, proxy alignment, or teacher supervision prevents them from learning compensation directly from incomplete samples, limiting their applicability to IMT. Other approaches admit incomplete samples and adapt shared

representations, prompts, or generated features to their available inputs Guo et al. (2024b); Zhang et al. (2025b). This raises a further question: how can the observed modality of one incomplete training sample help compensate for its absence in another?

2.2 Missing-Modality Compensation

Missing-modality compensation operates at the image or feature level. Image-level methods synthesize missing inputs before downstream processing, as exemplified by cycle-consistent translation Zhu et al. (2017), while MoPoE-VAE models multimodal observations in a latent probabilistic framework Sutter et al. (2021). Feature-level methods learn latent representations through incomplete multi-view learning, joint embedding, adaptation, debiasing, or masked modeling Gao & Pu (2025); Dai et al. (2025); Kim & Kim (2025); Reza et al. (2025); Do et al. (2025). ShaSpec separates shared and modality-specific representations Wang et al. (2023), whereas ActionMAE and M3AE reconstruct masked representations Woo et al. (2023); Liu et al. (2023). Many compensation methods infer the missing-modality representation from the available modality of the same sample. Although this preserves current-scene semantics, incomplete samples cannot directly supervise the required cross-modal mapping. Retrieval-based methods offer another source of information. RAGPT and REDEEM retrieve modality-complete instances from paired reference memories to support missing-modality compensation Lang et al. (2025a;b). In an IMT training set, this excludes the observed features of incomplete samples from the reference pool, leaving retrieval dependent on the remaining complete pairs. GenFuse instead uses features observed in both complete and incomplete training samples as cross-sample information and combines them with the current sample's semantics to construct the proxy.

3 Methodology

In this section, we formally define the problem of incomplete multimodal training and introduce our proposed Cross-Sample Genetic Fusion (GenFuse) framework.

3.1 Problem Setting and Overview

We consider incomplete multimodal training with two modalities, x_1 and x_2. Let $\mathcal{D} = \mathcal{D}_{T1} \cup \mathcal{D}_{T2} \cup \mathcal{D}_{T3}$ denote the labeled training set, where $\mathcal{D}_{T1}$ contains samples with only x_1, $\mathcal{D}_{T2}$ contains complete aligned pairs (x_1, x_2), and $\mathcal{D}_{T3}$ contains samples with only x_2. Modality-specific encoders E_m produce $f_m = E_m(x_m) \in \mathbb{R}^d$ whenever modality m is observed. For an incomplete sample with available modality m and missing modality m', where $m \neq m'$, GenFuse constructs $f_{m'}^{\text{proxy}}$ and fuses it with f_m for classification with genetic consistency supervision. Complete samples supply observed features for bidirectional proxy-alignment supervision.

As illustrated in Figure 2 for $(m, m') \in \{(1, 2), (2, 1)\}$, the available feature f_m serves as the semantic parent. GenFuse constructs a modality parent $\tilde{f}_{m'}$ by aggregating features retrieved from $\mathcal{B}_{m'}$, the feature bank corresponding to the missing modality. The Genetic Generation Module (GGM) then applies feature-wise crossover and range-scaled mutation to produce $f_{m'}^{\text{proxy}}$. GenFuse maintains a FIFO bank $\mathcal{B}_m$ of observed features per modality for cross-sample matching.

3.2 Cross-Sample Parental Matching (CSPM)

For an incomplete sample, the available feature f_m extracted from the observed modality of the current sample carries its scene semantics. However, it cannot by itself provide a representation of the missing modality. Conversely, a bank entry in $\mathcal{B}_{m'}$ provides a representation encoded from modality m', yet it may come from a training sample whose scene only partially matches the current one. Directly treating a single retrieved entry as the proxy would make it overly dependent on the sample-specific content of one cross-sample match. CSPM mitigates this dependence by constructing a modality parent from multiple retrieved features, with their aggregation conditioned on the semantic parent. Nevertheless, the resulting $\tilde{f}_{m'}$ remains composed entirely of bank features and is therefore not used directly as the proxy. GGM subsequently recombines $\tilde{f}_{m'}$ with f_m to generate the proxy.

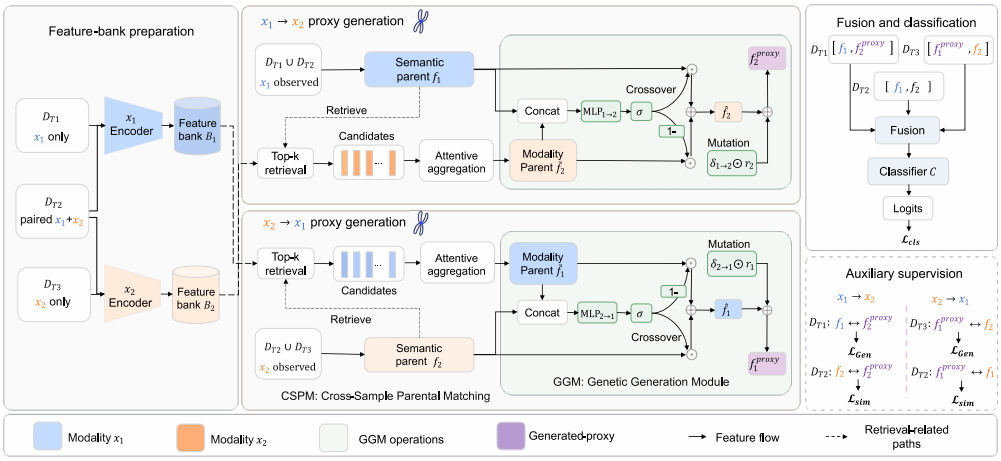


Figure 2: Overview of the proposed GenFuse architecture. CSPM constructs a modality parent by top-k retrieval and attention-weighted aggregation from the corresponding feature bank. GGM combines it with the observed semantic parent through feature-wise crossover and range-scaled mutation to generate a proxy for the missing modality. The diagram depicts both direct fusion of observed pairs and fusion with a generated proxy.

Cross-Sample Candidate Retrieval. For a generation direction with available modality m and missing modality m', let $b_{m',j} \in \mathbb{R}^d$ denote the j-th entry of the feature bank $\mathcal{B}_{m'}$. CSPM ranks the bank entries using the raw inner product $f_m^\top b_{m',j}$ and retains the top-k entries as coarse candidates:

$$\mathcal{I}_k = \text{TopK}_j \left(f_m^\top b_{m',j} \right), \tag{1}$$

where $\mathcal{I}_k$ contains the indices of the k bank entries with the largest matching scores.

Attentive Aggregation. A single retrieved candidate may contain sample-specific content that is only partially relevant to the current scene. CSPM therefore retains a set of high-scoring candidates rather than committing to one bank entry. Since the relevance of these candidates can still vary, CSPM conditions attention-weighted aggregation on the semantic parent f_m, allowing more compatible candidates to receive larger weights instead of relying on a single retrieval.

Specifically, after top-k selection, CSPM projects the semantic parent and each selected candidate into a query–key space:

$$Q = \mathbf{W}_Q f_m, \qquad K_j = \mathbf{W}_K b_{m',j}, \tag{2}$$

where $\mathbf{W}_Q$ and $\mathbf{W}_K$ are learnable projections. The attention weight for each selected entry is

$$a_j = \frac{\exp\left(Q^\top K_j / \sqrt{d}\right)}{\sum_{\ell \in \mathcal{I}_k} \exp\left(Q^\top K_\ell / \sqrt{d}\right)}. \tag{3}$$

The selected bank features serve directly as values, and their attention-weighted sum is ℓ_2-normalized to form the modality parent:

$$\tilde{f}_{m'} = \frac{\sum_{j \in \mathcal{I}_k} a_j b_{m',j}}{\left\| \sum_{j \in \mathcal{I}_k} a_j b_{m',j} \right\|_2}. \tag{4}$$

The projected query–key space is used only to reweight the selected candidates and does not alter the initial inner-product ranking. Figure 5(a–b) evaluates this coarse retrieval rule empirically. The resulting $\tilde{f}_{m'}$ serves as the modality parent, incorporating representations of the missing modality from the training set. CSPM outputs the parent pair $(f_m, \tilde{f}_{m'})$ to the GGM for subsequent proxy generation.

3.3 GENETIC GENERATION MODULE (GGM)

For an incomplete sample, neither parent alone provides all the information required to construct a proxy feature for the missing modality. The semantic parent carries the scene semantics of the current sample but is encoded from the available modality. In contrast, the modality parent incorporates representations of the missing modality from other training samples, but it may not fully match the current scene semantics. This complementarity motivates their genetic recombination in GGM. GGM performs this recombination through feature-wise crossover followed by range-scaled mutation.

Feature-wise Crossover. GGM first recombines the two parents through a direction-specific feature-wise gate. For direction $m \to m'$, the gate is computed from their concatenated features:

$$\mathbf{g}_{m\to m'} = \sigma\left(\mathrm{MLP}_{m\to m'}\left([f_m, \tilde{f}_{m'}]\right)\right), \tag{5}$$

where $\mathrm{MLP}_{m\to m'}$ is direction-specific, $[\cdot,\cdot]$ denotes concatenation, and σ is the sigmoid function. Thus, $\mathbf{g}_{m\to m'} \in (0,1)^d$ contains one gate value for each feature dimension. The crossover feature is then computed as

$$\hat{f}_{m'} = \mathbf{g}_{m\to m'} \odot f_m + (1 - \mathbf{g}_{m\to m'}) \odot \tilde{f}_{m'}, \tag{6}$$

where $\odot$ denotes element-wise multiplication.

This operation realizes crossover as feature-wise interpolation rather than a scalar-weighted combination of two vectors. Different feature dimensions can therefore receive different proportions of the two parents. Unlike conventional gated fusion between modalities observed in the same sample, GenFuse combines a semantic parent from the current sample with a modality parent constructed from the feature bank corresponding to the missing modality.

Range-scaled Mutation. Feature-wise crossover produces an interpolation between the two parent values in each feature dimension. Since each gate value lies between zero and one, crossover alone cannot introduce a direction-specific residual beyond this interpolation. GGM therefore applies range-scaled mutation to adjust the crossover feature according to the element-wise range of the corresponding feature bank:

$$r_{m'} = \max_{b\in\mathcal{B}_{m'}} b - \min_{b\in\mathcal{B}_{m'}} b. \tag{7}$$

The proxy is generated by

$$f_{m'}^{\mathrm{proxy}} = \frac{\hat{f}_{m'} + \delta_{m\to m'} \odot r_{m'}}{\left\|\hat{f}_{m'} + \delta_{m\to m'} \odot r_{m'}\right\|_2}, \tag{8}$$

where $\delta_{m\to m'} \in \mathbb{R}^d$ is a direction-specific learnable vector shared across samples. The mutation is implemented as a learnable offset whose per-dimension magnitude is scaled by the variation of the corresponding feature dimension in $\mathcal{B}_{m'}$. Feature-wise crossover recombines the two parents, while range-scaled mutation adds a direction-specific correction beyond pure interpolation. The resulting proxy $f_{m'}^{\mathrm{proxy}}$ is fused with the available feature f_m for classification.

3.4 TRAINING OBJECTIVES

GenFuse applies supervision according to modality availability during training. For incomplete samples, the missing-modality feature is unavailable, so the available feature provides semantic-consistency supervision for proxy generation. For complete samples, the corresponding observed features provide bidirectional proxy-alignment supervision. Both sample types contribute to the classification objective.

Classification Objective. Depending on the training path, the fusion module receives either an observed feature pair (f_1, f_2) or an available feature paired with a generated proxy. The two features are concatenated and passed to a two-layer MLP classifier. The classification loss $\mathcal{L}_{\mathrm{cls}}$ is computed as standard cross-entropy over the samples in the corresponding training path. Throughout this subsection, $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity.

Genetic Consistency Loss for Incomplete Samples. For incomplete samples, the feature of the missing modality is unavailable and therefore cannot provide direct proxy-alignment supervision. We introduce the genetic consistency loss $\mathcal{L}_{\mathrm{Gen}}$ to encourage the generated proxy to preserve the scene semantics conveyed by the available feature:

$$\ell_{\mathrm{Gen}}(\mathcal{S}; u, v) = \begin{cases} \dfrac{1}{|\mathcal{S}|} \displaystyle\sum_{i \in \mathcal{S}} \left[1 - \mathrm{sim}(u_i, v_i)\right], & |\mathcal{S}| > 0, \\ 0, & |\mathcal{S}| = 0, \end{cases} \tag{9}$$

$$\begin{aligned} \mathcal{L}_{\mathrm{Gen}} = \frac{1}{2}\Big[& \ell_{\mathrm{Gen}}(\mathcal{D}_{T1}; f_2^{\mathrm{proxy}}, f_1) \\ & + \ell_{\mathrm{Gen}}(\mathcal{D}_{T3}; f_1^{\mathrm{proxy}}, f_2)\Big]. \end{aligned} \tag{10}$$

The empty-set definition disables the corresponding direction when no such incomplete samples are present. When both incomplete subsets are empty, $\mathcal{L}_{\mathrm{Gen}} = 0$. This objective provides semantic supervision for proxy generation without requiring an observed feature from the missing modality. It does not enforce feature-level reconstruction.

Proxy Similarity Loss for Complete Samples. For complete samples, both modality features are observed, allowing each generated proxy to be aligned with the corresponding observed feature. GenFuse alternately treats each modality as the available modality and generates a proxy for the other. Let $f_{m',i}^{\mathrm{proxy}}$ denote the proxy generated for modality m' of sample i from the other observed modality. The proxy similarity loss is defined as

$$\mathcal{L}_{\mathrm{sim}} = \frac{1}{2|\mathcal{D}_{T2}|} \sum_{i \in \mathcal{D}_{T2}} \sum_{m' \in \{1,2\}} \left[1 - \mathrm{sim}\left(f_{m',i}^{\mathrm{proxy}}, f_{m',i}\right)\right], \tag{11}$$

This loss is evaluated in both generation directions on $\mathcal{D}_{T2}$. Complete samples can provide bidirectional proxy-similarity supervision but do not contribute to the genetic consistency loss.

Overall Objective. The combined objective is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{cls}} + \lambda_2 \mathcal{L}_{\mathrm{Gen}} + \lambda_3 \mathcal{L}_{\mathrm{sim}}. \tag{12}$$

The auxiliary losses are applied according to supervision availability: $\mathcal{L}_{\mathrm{sim}}$ is evaluated only on complete samples, whereas $\mathcal{L}_{\mathrm{Gen}}$ is evaluated only on incomplete samples. The classification loss is computed for both types. The coefficients λ_1, λ_2, and λ_3 balance the classification, genetic-consistency, and proxy-alignment objectives. The optimization details are in Appendix A.

Stage-wise Training. Stage 1 enqueues detached observed training features before retrieval. It classifies complete samples through direct fusion and incomplete samples through proxy fusion, optimizing the combined objective in Eq. 12. Stage 2 keeps the banks fixed but continues updating the encoders, CSPM, GGM, and classifier using only $\mathcal{L}_{\mathrm{cls}}$ in the reported IMT experiments. With a designated missing direction, complete samples also follow the proxy classification path; with mixed availability patterns, they retain direct fusion. The banks remain fixed at inference, and test samples are never inserted.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate GenFuse on Houston Debes et al. (2014) (144-band HSI/LiDAR DSM; 15,029 samples from 15 urban classes), Augsburg Baumgartner et al. (2012) (78,294 samples from seven land-cover classes), and Trento Rasti et al. (2017) (co-registered HSI/LiDAR DSM; 30,214 samples from six classes). All Augsburg experiments use only HSI–SAR. For each dataset, 40 labeled pixels per class are randomly assigned to training, and the remaining labeled pixels form the test set; spatial separation is not enforced.

Table 1: Specific modality-missing results (%). η_{train} is the fraction of incomplete training samples; "W/o" denotes the modality removed at test time. Best and second-best values are highlighted and underlined.

Missing Rate	Method	Houston						Augsburg						Trento					
		W/o HSI			W/o LiDAR			W/o HSI			W/o SAR			W/o HSI			W/o LiDAR		
		OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ	OA	AA	κ
0%	ShaSpec	49.4	50.4	45.6	94.2	95.2	93.8	48.0	32.6	33.1	86.4	74.0	80.9	87.8	87.8	84.0	96.1	96.0	94.8
	DGDNet	59.1	60.4	55.9	93.5	94.5	93.0	51.9	30.7	31.7	78.4	74.7	71.1	93.5	91.7	91.4	97.6	96.9	96.8
	LDS2AE	65.1	66.6	62.4	96.0	96.8	95.7	47.7	29.5	17.6	85.9	**84.4**	80.8	88.5	87.7	84.9	98.2	**97.6**	97.6
	MCF-Net	34.2	31.3	28.5	72.5	70.3	70.2	37.0	22.1	26.3	79.5	70.0	72.3	65.2	33.3	47.9	85.3	77.7	80.2
	DPMamba	61.5	64.7	58.5	97.0	**97.6**	**96.8**	79.7	**62.5**	**72.5**	88.3	**84.4**	83.9	94.4	93.3	92.6	98.3	97.1	97.8
	GenFuse	**67.4**	**70.6**	**64.9**	**97.1**	**97.6**	**96.8**	**80.8**	47.4	72.3	**89.2**	84.0	**85.0**	**94.7**	**93.9**	**92.9**	**98.5**	97.1	**98.1**
30%	ShaSpec	34.2	29.1	28.4	65.8	69.3	63.0	31.8	33.9	21.3	79.2	67.5	71.6	88.2	84.1	84.5	95.3	92.6	93.7
	DGDNet	46.0	46.0	41.8	31.6	37.7	26.4	39.4	21.9	15.9	80.9	63.3	74.2	91.5	86.3	88.1	96.6	95.2	95.5
	LDS2AE	56.2	58.3	52.8	34.9	42.8	29.6	28.6	28.8	13.9	80.7	74.7	74.1	86.8	87.0	82.7	98.4	97.3	97.8
	MCF-Net	47.9	49.8	44.1	36.8	43.9	31.6	34.9	15.8	11.9	79.3	70.0	72.2	93.0	83.7	90.6	97.5	95.8	96.7
	DPMamba	58.4	60.3	55.1	43.5	51.4	39.4	50.2	32.0	33.2	84.8	74.7	79.3	94.8	91.8	93.0	96.9	95.3	95.9
	GenFuse	**66.4**	**68.7**	**63.8**	**95.7**	**96.5**	**95.3**	**59.1**	**56.2**	**49.1**	**87.0**	**80.0**	**82.1**	**95.2**	**92.8**	**93.6**	**99.0**	**97.8**	**98.6**
70%	ShaSpec	58.0	63.0	54.5	38.2	35.7	33.3	44.7	32.2	32.1	85.5	72.0	80.0	86.6	78.2	82.1	93.1	89.2	90.8
	DGDNet	43.5	43.5	39.3	32.6	37.7	26.9	43.6	37.9	32.9	80.5	67.0	73.3	91.7	86.2	88.9	95.5	93.2	94.0
	LDS2AE	53.0	55.3	49.5	16.6	19.5	9.4	41.1	27.1	19.6	81.5	76.0	75.0	88.0	86.9	84.2	97.4	96.1	96.6
	MCF-Net	46.9	47.5	42.9	36.3	43.2	31.5	26.6	23.2	3.4	57.2	35.2	39.5	95.5	**93.6**	93.9	97.6	94.1	96.8
	DPMamba	54.0	55.7	50.5	56.2	63.2	52.7	37.2	24.9	15.1	71.7	68.9	63.8	**95.8**	88.7	**94.4**	97.3	94.5	96.3
	GenFuse	**64.9**	**66.6**	**62.2**	**92.9**	**94.4**	**92.4**	**70.9**	**54.3**	**61.6**	**86.2**	**80.2**	**81.1**	95.4	92.5	93.8	**98.6**	**97.6**	**98.1**

Implementation Details We implement GenFuse in PyTorch on one NVIDIA GeForce RTX 4090. Each modality bank stores 16,384 64-dimensional features. Stage 1/2 run for 600/300 epochs with AdamW (10^{-4}, weight decay 10^{-4})/Adam (10^{-4}), respectively, and cosine annealing with warm restarts; the banks are frozen in Stage 2. For Stage 1, the loss weights are $\lambda_1 = 1$, $\lambda_2 = 0.005$, and $\lambda_3 = 0.01$. In specific experiments, Stage 1 uses $k = 32$; Stage 2 uses $k = 32$ for HSI-missing inputs and $k = 4$ for LiDAR/SAR-missing inputs. On Trento, the two banks occupy 8.0 MiB; Stage 1 takes 202.0 ms/batch (77.1 without retrieval), and missing-input inference takes 0.162 ms/sample. See Appendix A for configurations and Appendix D for profiling.

Modality-Missing Simulation. During training, sample-wise missing masks are assigned at rate η_{train} using a fixed random seed. Each incomplete sample retains one uniformly selected modality, while the other is replaced by an all-zero tensor; the remaining samples retain both modalities. We evaluate two protocols. In the *specific* protocol, one specified modality is deterministically removed from every test sample, so 0%, 30%, and 70% refer only to the proportions of incomplete training samples. In the *probabilistic* protocol, training and test samples follow the same masking rule with $\eta_{\text{train}} = \eta_{\text{test}} \in \{30\%, 70\%\}$; the resulting test set contains both complete and single-modality-missing samples. All methods use identical data splits and missing protocols. The main tables report single-run results; the paired three-seed check is reported separately below.

4.2 Experimental Results

4.2.1 Comparison with Other Methods

We compare GenFuse with five incomplete multimodal classification methods: ShaSpec Wang et al. (2023), DGDNet Wei et al. (2023b), LDS2AE Qu et al. (2024a), MCF-Net Song et al. (2025), and DPMamba Yang et al. (2025). We report overall accuracy (OA), average accuracy (AA), and Kappa (κ) under identical splits and missing protocols.

Specific Modality Missing. Table 1 reports results when a specified modality is removed at test time. At $\eta_{\text{train}} = 30\%$, GenFuse ranks first on every reported metric across all three datasets. At 70%, it reaches 92.9% OA on Houston without LiDAR and 70.9% on Augsburg without HSI, exceeding the respective runners-up by 36.7 and 26.2 pp. At 0% on Augsburg without HSI, its OA slightly exceeds DPMamba's (80.8% versus 79.7%), but its AA is lower (47.4% versus 62.5%). On

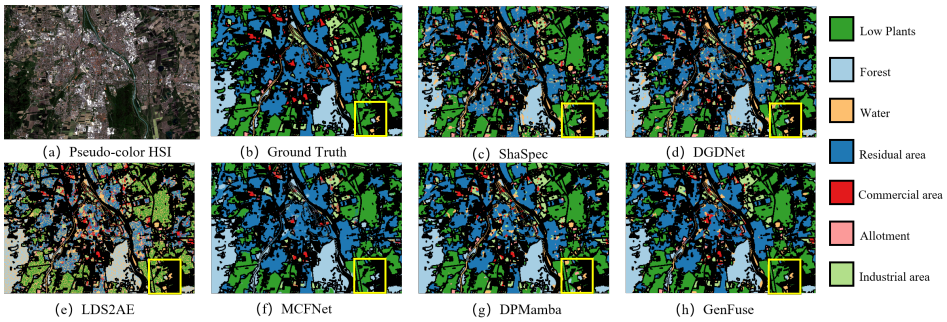


Figure 3: Augsburg classification maps with SAR missing at test time ($\eta_{\text{train}} = 0\%$, specific protocol). Highlighted regions show prediction differences.

Trento, the margins narrow: at 70% without HSI, DPMamba exceeds GenFuse in OA (95.8% versus 95.4%), whereas GenFuse leads when LiDAR is missing.

Robustness Checks on Houston. For Houston with 70% incomplete training samples and LiDAR missing at test time, GenFuse achieves $92.1 \pm 0.7\%$ OA across three runs, compared with $38.9 \pm 18.4\%$ for DPMamba (mean $\pm$ sample SD). GenFuse has higher OA in all three matched runs. Using the same three seeds, a paired control excluding all same-ID bank entries yields $93.0 \pm 1.0\%$ OA for GenFuse, versus $92.1 \pm 0.7\%$ without exclusion. In this condition, same-ID entries arise only during complete-sample proxy alignment. See Appendix B for run-level results.

Table 2: Probabilistic missing results (%) at $\eta_{\text{train}} = \eta_{\text{test}}$; test sets mix complete and single-modality-missing samples. Best and second-best values are highlighted and underlined.

$\eta_{\text{train}} = \eta_{\text{test}}$	Method	Houston (HSI/LiDAR)			Augsburg (HSI/SAR)			Trento (HSI/LiDAR)		
		OA	AA	κ	OA	AA	κ	OA	AA	κ
30%	ShaSpec	62.01	63.40	58.90	72.14	63.00	63.14	96.56	94.75	95.42
	DGDNet	55.34	57.37	51.73	<u>75.61</u>	54.73	67.49	97.17	95.00	96.21
	LDS2AE	62.15	65.32	59.11	75.58	<u>69.75</u>	<u>67.58</u>	95.46	93.55	93.95
	MCF-Net	<u>64.07</u>	<u>68.44</u>	<u>61.30</u>	60.15	35.79	41.60	<u>98.25</u>	<u>96.16</u>	<u>97.66</u>
	DPMamba	55.31	61.34	52.32	72.31	66.84	64.12	97.04	95.20	96.04
	GenFuse	**89.51**	**90.67**	**88.66**	**79.84**	**71.71**	**72.50**	**98.31**	**97.34**	**97.73**
70%	ShaSpec	<u>56.33</u>	<u>57.80</u>	<u>52.86</u>	<u>67.57</u>	<u>59.67</u>	<u>57.72</u>	93.12	90.09	90.86
	DGDNet	40.50	43.74	35.97	54.03	50.47	43.63	95.36	91.35	93.81
	LDS2AE	42.67	44.60	38.05	58.02	58.65	48.31	94.34	92.90	92.47
	MCF-Net	48.35	52.25	44.28	49.85	25.99	24.69	<u>97.40</u>	94.98	<u>96.52</u>
	DPMamba	47.00	53.06	43.00	58.78	50.47	45.01	**97.94**	<u>95.49</u>	**97.24**
	GenFuse	**88.13**	**89.44**	**87.18**	**78.00**	**69.74**	**70.48**	97.34	**96.16**	96.44

Probabilistic Modality Missing. At 30% missingness, GenFuse ranks first in OA, AA, and κ on all three datasets, although its Trento OA is close to MCF-Net's (Table 2). At 70%, its OA exceeds the runners-up by 31.80 pp on Houston and 10.43 pp on Augsburg. On Trento, DPMamba leads in OA and κ, while GenFuse leads in AA and trails by 0.60 pp in OA. Thus, GenFuse maintains competitive performance under both directional and probabilistic modality missing, with the largest gains on Houston and Augsburg.

Visual Comparison. Figure 3 compares classification maps on Augsburg. Baseline maps contain spatially scattered misclassified pixels and less coherent predictions within otherwise contiguous land-cover regions. GenFuse produces more spatially coherent regions and cleaner class boundaries, consistent with its quantitative performance on this dataset.

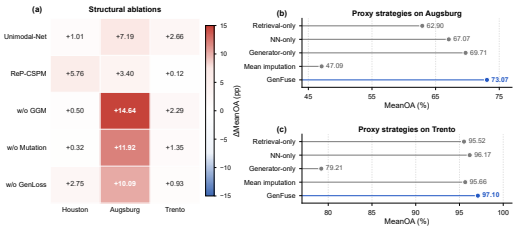


Figure 4: Ablation and proxy-strategy results at $\eta_{\text{train}} = 30\%$. (a) MeanOA decreases from GenFuse; (b–c) proxy alternatives on Augsburg and Trento.

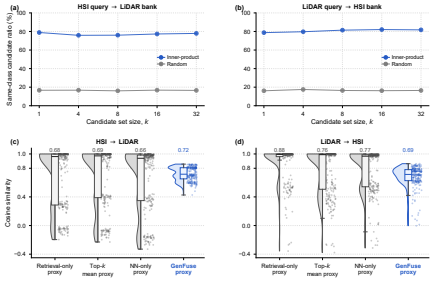


Figure 5: Trento retrieval and proxy analysis: (a–b) Same-class ratios for inner-product and random retrieval; (c–d) proxy-to-observed cosine similarity.

4.3 Component and Mechanism Analysis

Structural Ablations. Figure 4(a) reports MeanOA drops averaged over two missing directions at $\eta_{\text{train}} = 30\%$. Unimodal-Net trains and evaluates a classifier using only the available modality; it trails GenFuse by 1.01 pp on Houston and 7.19 pp on Augsburg. On Augsburg, removing GGM, mutation, or $\mathcal{L}_{\text{Gen}}$ reduces MeanOA by 14.64, 11.92, and 10.09 pp, respectively; replacing CSPM with nearest-feature selection costs 5.76 pp on Houston. All Trento variants exceed 94% MeanOA and differ from GenFuse by 0.12–2.66 pp. Details appear in Appendix B.

Alternative Proxy Strategies. The four alternatives are independently trained classification variants. Retrieval-only uses the attention-weighted modality parent as the proxy; NN-only uses the highest-scoring bank feature; Generator-only combines the semantic parent with a learned proxy vector without retrieval; and Mean imputation uses the running mean of observed training features. These variants are trained independently. In Figure 4(b–c), GenFuse leads the strongest alternative on Augsburg (73.07% versus 69.71% MeanOA) and Trento (97.10% versus 96.17%). These results show that neither direct retrieval nor generation without retrieved parents consistently matches their combination in GenFuse. The larger separation on Augsburg indicates a stronger benefit in the more challenging HSI–SAR setting.

Retrieval Validity and Proxy Alignment. Using a fixed pool of re-encoded training features, Figure 5(a–b) reports same-class ratios for raw-inner-product versus random top-k retrieval. Raw retrieval exceeds random by 59.13–65.86 pp after averaging over models trained with 30%, 50%, and 70% missing samples. Labels and observed counterparts are used only for evaluation, not for candidate selection or proxy construction. In Figure 5(c–d), GenFuse's mean proxy-to-observed cosine similarity is 0.72 for HSI-to-LiDAR and 0.69 in reverse; Retrieval-only reaches 0.88 in reverse despite lower OA. Together with Figure 4(b–c), these results show that a higher proxy-to-observed-feature cosine value does not necessarily correspond to higher OA. Taken together, Figures 4 and 5 show that the gains arise not from retrieval alone but from recombining retrieved missing-modality representations with the current sample's semantic parent.

5 Conclusion

This work presented GenFuse for incomplete multimodal training in remote sensing. GenFuse retrieves representations of the missing modality to form a modality parent and recombines it with the current sample's semantic parent through feature-wise crossover and range-scaled mutation. The resulting proxy complements the available feature for classification. Experiments on three datasets under specific and probabilistic modality-missing protocols showed that GenFuse achieved the best results in most evaluated conditions, while component and proxy-strategy comparisons supported the value of the full design over simpler alternatives. The present evaluation is limited to two-modality classification under controlled masking. Future work will examine extensions to more than two modalities and structured missingness caused by coverage gaps or sensor degradation.

AI USE STATEMENT

The authors drafted the manuscript and developed the GenFuse method. Generative AI tools assisted with language editing, identifying related literature, and reviewing experimental protocols, scripts, and results. The authors verified AI-assisted material against the cited sources and experimental records and take responsibility for the final paper.

REFERENCES

Andreas Baumgartner, Peter Gege, Claas Köhler, Karim Lenhard, and Thomas Schwarzmaier. Characterisation methods for the hyperspectral sensor hyspex at dlr's calibration home base. *Proceedings of SPIE - The International Society for Optical Engineering*, 8533, 09 2012. doi: 10.1117/12.974664.

Jiajun Chen, Sai Cheng, Yuan Yutao, YiRui Zhang, Haitao Yuan, Peng Peng, and Yi Zhong. Promise: Prompt-attentive hierarchical contrastive learning for robust cross-modal representation with missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (24):20076–20082, Mar. 2026. doi: 10.1609/aaai.v40i24.39093. URL https://ojs.aaai.org/index.php/AAAI/article/view/39093.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24507–24517, October 2025.

Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, Wilfried Philips, Saurabh Prasad, Qian Du, and Fabio Pacifici. Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2405–2418, 2014. doi: 10.1109/JSTARS.2014.2305441.

Minh Kha Do, Kang Han, Phu Lai, Khoa T. Phan, and Wei Xiang. Robsense: A robust multi-modal foundation model for remote sensing with static, temporal, and incomplete data adaptability. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7427–7436, 2025. doi: 10.1109/CVPR52734.2025.00696.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Stil: Semi-supervised tabular-image learning for comprehensive task-relevant information exploration in multimodal classification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15549–15559, 2025. doi: 10.1109/CVPR52734.2025.01449.

Siyi Du, Xinzhe Luo, Declan O'regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PWhDUWRVhM.

Xin Gao and Jian Pu. Deep incomplete multi-view learning via cyclic permutation of VAEs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s4MwstmB8o.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27662–27673, 2024a. doi: 10.1109/CVPR52733.2024.02613.

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25015–25025, 2025. doi: 10.1109/CVPR52734.2025.02329.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1726–1736, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.94. URL https://aclanthology.org/2024.acl-long.94/.

Wenqi Han, Jie Geng, Zhengyi Xu, and Wen Jiang. Multimodal heterogeneous hypergraph learning for incomplete multimodal semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. doi: 10.1109/TGRS.2025.3604075.

Ling Huang, Wenqian Dong, Song Xiao, Jiahui Qu, Yuanbo Yang, and Yunsong Li. Language-guided visual prompt compensation for multi-modal remote sensing image classification with modality absence. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 5161–5170, New York, NY, USA, 2024a. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681563. URL https://doi.org/10.1145/3664647.3681563.

Xin Huang, Aigerim Rymbekova, Olga Dolgova, Oscar Lao, and Martin Kuhlwilm. Harnessing deep learning for population genetic inference. *Nature Reviews Genetics*, 25(1):61–78, 2024b.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873, June 2025.

Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 171–187, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73016-0.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):18035–18043, 2025a. doi: 10.1609/aaai.v39i17.33984.

Jian Lang, Rongpei Hong, Zhangtao Cheng, Ting Zhong, Yong Wang, and Fan Zhou. REDEEMing modality information loss: Retrieval-guided conditional generation for severely modality missing learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 1241–1252. Association for Computing Machinery, 2025b. doi: 10.1145/3711896.3737101.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, 2023. doi: 10.1109/CVPR52729.2023.01435.

Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1657–1665, Jun. 2023. doi: 10.1609/aaai.v37i2.25253. URL https://ojs.aaai.org/index.php/AAAI/article/view/25253.

Xiao Liu, Fei Jin, Shuxiang Wang, Jie Rui, Xibing Zuo, Xiaobing Yang, and Chuanxiang Cheng. Multimodal online knowledge distillation framework for land use/cover classification using full or missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. doi: 10.1109/TGRS.2024.3388604.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18156–18165, 2022. doi: 10.1109/CVPR52688.2022.01764.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3215–3224, October 2025.

Jiahui Qu, Yuanbo Yang, Wenqian Dong, and Yufei Yang. Lds2ae: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14731–14739, Mar. 2024a. doi: 10.1609/aaai.v38i13.29391. URL https://ojs.aaai.org/index.php/AAAI/article/view/29391.

Jiahui Qu, Lijian Zhang, Wenqian Dong, Nan Li, and Yunsong Li. Shared-private decoupling-based multilevel feature alignment semisupervised learning for hsi and lidar classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024b. doi: 10.1109/TGRS.2024.3492499.

Behnood Rasti, Pedram Ghamisi, and Richard Gloaguen. Hyperspectral and lidar fusion using extinction profiles and total variation component analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3997–4007, 2017. doi: 10.1109/TGRS.2017.2686450.

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, 2025. doi: 10.1109/TPAMI.2024.3476487.

Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Ogp-net: Optical guidance meets pixel-level contrastive distillation for robust multi-modal and missing modality segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):6922–6930, Apr. 2025. doi: 10.1609/aaai.v39i7.32743. URL https://ojs.aaai.org/index.php/AAAI/article/view/32743.

Qiya Song, Feng Mo, Kexin Ding, Lin Xiao, Renwei Dian, Xudong Kang, and Shutao Li. Mcfnet: Multiscale cross-domain fusion network for hsi and lidar data joint classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–12, 2025. doi: 10.1109/TGRS.2025.3567297.

Haoqin Sun, Xugang Lu, Jingguang Tian, Jiaming Zhou, Jiabei He, Hui Wang, Xiangyu Kong, Xinhui Hu, and Yong Qin. Progressive learning framework with missing modality reconstruction for multimodal emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 34:29–43, 2026. doi: 10.1109/TASLPRO.2025.3640882.

Thomas M. Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal ELBO. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=5Y21V0RDBV.

Sirnam Swetha, Mamshad Nayeem Rizve, Nina Shvetsova, Hilde Kuehne, and Mubarak Shah. Preserving modality structure improves multi-modal learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21936–21946, 2023. doi: 10.1109/ICCV51070.2023.02010.

Haoyu Wang, Xiaomin Liu, Zhenzhuang Qiao, Guoqing Wang, and Haotian Chen. Multimodal remote sensing data classification based on gaussian mixture variational dynamic fusion network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3394462.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887, 2023. doi: 10.1109/CVPR52729.2023.01524.

Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20039–20049, 2023a. doi: 10.1109/CVPR52729.2023.01919.

Shicai Wei, Yang Luo, and Chunbo Luo. Diversity-guided distillation with modality-center regularization for robust multimodal remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023b. doi: 10.1109/TGRS.2023.3336297.

Shicai Wei, Yang Luo, Xiaoguang Ma, Peng Ren, and Chunbo Luo. Msh-net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023c. doi: 10.1109/TGRS.2023.3265650.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i3.25378. URL https://doi.org/10.1609/aaai.v37i3.25378.

Xiaomeng Xin, Heping Song, and Jianping Gou. A new similarity-based relational knowledge distillation method. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3535–3539, 2024. doi: 10.1109/ICASSP48485.2024.10447596.

Aijun Xu, Zhaohui Xue, and Hao Wu. How does missing-modality affect cross-domain remote sensing image classification: A synergistic inter-modal proxy fine-tuning solution. *ISPRS Journal of Photogrammetry and Remote Sensing*, 235:153–170, 2026. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2026.02.040. URL https://www.sciencedirect.com/science/article/pii/S0924271626000997.

Yueguang Yang, Jiahui Qu, Ling Huang, and Wenqian Dong. Dpmamba: distillation prompt mamba for multimodal remote sensing image classification with missing modalities. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/248. URL https://doi.org/10.24963/ijcai.2025/248.

Burak Yelmen and Flora Jay. An overview of deep generative models in functional and evolutionary genomics. *Annual Review of Biomedical Data Science*, 6(1):173–189, 2023.

Limeng Zhang, Zenghui Zhang, Tao Zhang, Gui Gao, and Wenxian Yu. Cdprompt: Multimodal change detection with in-domain prompt in missing modality scenarios. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–18, 2025a. doi: 10.1109/TGRS.2025.3598803.

Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1881–1890, October 2025b.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8048–8059, 2023. doi: 10.1109/ICCV51070.2023.00742.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25912–25921, 2024. doi: 10.1109/CVPR52733.2024.02448.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.

APPENDIX

A EXPERIMENTAL PROTOCOLS AND IMPLEMENTATION

A.1 DATASETS, SPLITS, AND CONFIGURATIONS

For every dataset, 40 labeled samples per class form the training split and all remaining labeled samples form the test split. Missingness is applied after this split, and every compared method uses the identical split and mask realization. Houston and Trento use HSI and LiDAR, whereas Augsburg uses HSI and SAR. Tables 3 and 4 report the stage-, direction-, and protocol-specific configurations.

Table 3: Detailed hyper-parameters for the specific modality-missing protocol. η_{train} is applied only during training, and "W/o" identifies the modality deterministically removed from every test sample. All rows use 64-dimensional features.

Dataset	Stage	Condition	η_{train}	Bank size	Top-k	Epochs	Patch size	Batch size
Houston (HSI/LiDAR)	Stage 1	Sample-wise missing	$\{0, 30, 70\}\%$	16,384	32	600	11	280
	Stage 2	W/o HSI	$\{0, 30, 70\}\%$	16,384	32	300	11	280
	Stage 2	W/o LiDAR	$\{0, 30, 70\}\%$	16,384	4	300	11	140
Augsburg (HSI/SAR)	Stage 1	Sample-wise missing	$\{0, 30, 70\}\%$	16,384	32	600	11	280
	Stage 2	W/o HSI	$\{0, 30, 70\}\%$	16,384	32	300	11	280
	Stage 2	W/o SAR	$\{0, 30, 70\}\%$	16,384	4	300	11	140
Trento (HSI/LiDAR)	Stage 1	Sample-wise missing	$\{0, 30, 70\}\%$	16,384	32	600	15	120
	Stage 2	W/o HSI	$\{0, 30, 70\}\%$	16,384	32	300	15	120
	Stage 2	W/o LiDAR	$\{0, 30, 70\}\%$	16,384	4	300	15	120

Table 4: Detailed hyper-parameters for the probabilistic modality-missing protocol corresponding to Table 2 of the main paper. The same sample-wise masking rule is used in training and testing, so test inputs include complete and single-modality-missing samples. All rows use 64-dimensional features.

Dataset	Stage	Masking rule	$\eta_{\text{train}} = \eta_{\text{test}}$	Bank size	Top-k	Epochs	Patch size	Batch size
Houston (HSI/LiDAR)	Stage 1	Sample-wise probabilistic	$\{30, 70\}\%$	16,384	32	600	11	280
	Stage 2	Same probabilistic rule	$\{30, 70\}\%$	16,384	8	300	11	280
Augsburg (HSI/SAR)	Stage 1	Sample-wise probabilistic	$\{30, 70\}\%$	16,384	32	600	9	140
	Stage 2	Same probabilistic rule	$\{30, 70\}\%$	16,384	8	300	9	140
Trento (HSI/LiDAR)	Stage 1	Sample-wise probabilistic	$\{30, 70\}\%$	16,384	32	600	15	120
	Stage 2	Same probabilistic rule	$\{30, 70\}\%$	16,384	8	300	15	120

Stage 1 uses AdamW with an initial learning rate of 10^{-4} and weight decay 10^{-4} for 600 epochs. Stage 2 uses Adam with an initial learning rate of 10^{-4} for 300 epochs. Both stages use cosine annealing with warm restarts. Each modality encoder contains two 1×1 reduction layers, four Transformer blocks with eight attention heads, and a 64-dimensional nonlinear projection head.

Tables 3 and 4 separate nominal epoch counts from stage- and direction-specific retrieval settings. The 16,384-entry banks and 64-dimensional features are fixed throughout. Under the specific protocol, Stage 2 uses the common direction-level rule $k = 32$ for HSI-missing inputs and $k = 4$ for LiDAR/SAR-missing inputs; under the probabilistic protocol, Stage 2 uses $k = 8$ for all three datasets. The tables also preserve the dataset-specific input patch and batch sizes instead of presenting one global setting.

A.2 MISSING-MODALITY PROTOCOLS

During training, sample-wise masks are assigned once using a fixed random seed. With probability η_{train}, a sample is incomplete: one of its two modalities is retained uniformly at random and the

Table 5: Objectives and bank states in the two-stage training protocol. $\mathcal{L}_{\mathrm{Gen}}$ contributes only for incomplete samples and $\mathcal{L}_{\mathrm{sim}}$ only for complete samples.

Protocol	Stage	Objectives	Bank state / evaluation
Specific	1	$\mathcal{L}_{\mathrm{cls}} + 0.005\mathcal{L}_{\mathrm{Gen}} + 0.01\mathcal{L}_{\mathrm{sim}}$	FIFO bank updated from observed training features
Specific	2	$\mathcal{L}_{\mathrm{cls}}$	Bank fixed; one modality removed from every test sample
Prob.-C1	1	$\mathcal{L}_{\mathrm{cls}} + 0.005\mathcal{L}_{\mathrm{Gen}} + 0.01\mathcal{L}_{\mathrm{sim}}$	Complete-modality reference
Prob.-C1	2	$\mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{sim}}$	Bank fixed; complete test inputs
Prob.-C2	1	$\mathcal{L}_{\mathrm{cls}} + 0.005\mathcal{L}_{\mathrm{Gen}} + 0.01\mathcal{L}_{\mathrm{sim}}$	FIFO bank updated from observed training features
Prob.-C2	2	$\mathcal{L}_{\mathrm{cls}}$	Bank fixed; $\eta_{\mathrm{train}} = \eta_{\mathrm{test}}$

other is replaced by an all-zero tensor of matching dimensions. The remaining samples retain both modalities.

Under the *specific modality-missing* protocol, $\eta_{\mathrm{train}} \in \{0\%, 30\%, 70\%\}$ denotes the proportion of incomplete *training* samples. At test time, one specified modality is deterministically removed from every test sample. Under the *probabilistic modality-missing* protocol, Condition 1 is a complete-modality reference that uses complete test inputs and is not reported in Table 2 of the main paper. Condition 2 is the IMT setting in Table 2: training and test samples use the same sample-wise masking rule with $\eta_{\mathrm{train}} = \eta_{\mathrm{test}} \in \{30\%, 70\%\}$, so the test set contains complete and single-modality-missing samples. All baselines use the same data splits and the same protocol-specific masks.

Table 5 makes the supervision boundary explicit. Complete samples provide the observed counterpart needed by $\mathcal{L}_{\mathrm{sim}}$, while incomplete samples contribute to classification and genetic consistency without exposing the missing-modality feature. Stage 2 retains trainable encoders and fusion modules but does not update the bank state.

A.3 Feature-Bank Lifecycle and Baseline Adaptation

GenFuse maintains one fixed-capacity FIFO bank for each modality. In Stage 1, detached encoder features from observed training modalities are enqueued before retrieval. Thus, a complete sample updates both banks, whereas an incomplete sample updates only the bank of its available modality; no feature is inserted for its missing modality. Stage 2 initializes from the Stage 1 bank state and keeps both banks fixed while the encoders, CSPM, GGM, fusion layers, and classifier remain trainable. At inference, all model parameters and bank states are fixed, and test samples never enter either bank.

The complete paired subset provides bidirectional proxy-similarity supervision: each modality can be treated as available while the corresponding observed feature is used by $\mathcal{L}_{\mathrm{sim}}$. Incomplete samples do not provide an observed feature of their missing modality. They instead use the available feature and generated proxy on the classification path and contribute to $\mathcal{L}_{\mathrm{Gen}}$. Therefore, GenFuse does not require missing-modality supervision for incomplete samples, but it does use the complete paired subset for proxy-alignment supervision.

All baselines are optimized on the same labeled split and mask realization. When a baseline expects two input tensors, the unavailable modality is supplied as the same all-zero tensor convention used for GenFuse. No test feature is used for model training, bank construction, or the analyses below.

B Complete Component and Proxy-Strategy Results

B.1 Structural Variants

The structural study uses the specific modality-missing protocol with $\eta_{\mathrm{train}} = 30\%$. Each reported MeanOA, MeanAA, or Meanκ is the arithmetic mean over the two deterministic test directions of the same dataset; it is not an average over random seeds. **Unimodal-Net** trains and tests a single-modality classifier. **ReP-CSPM** replaces top-k retrieval and attentive aggregation with a single nearest bank feature while retaining the downstream generation path. **w/o GGM** directly fuses the

Table 6: Complete structural-ablation results (%) at $\eta_{\text{train}} = 30\%$ under the specific modality-missing protocol. Each metric is averaged over the two deterministic test directions.

Method	Houston			Augsburg			Trento		
	MeanOA	MeanAA	Meanκ	MeanOA	MeanAA	Meanκ	MeanOA	MeanAA	Meanκ
Unimodal-Net	80.02	82.27	78.48	65.88	54.48	54.06	94.44	88.72	92.57
ReP-CSPM	75.27	72.99	73.22	69.67	60.01	58.30	96.98	95.09	95.97
w/o GGM	80.53	82.09	79.01	58.43	55.90	50.11	94.81	92.98	93.10
w/o Mutation	80.71	82.30	79.21	61.15	57.27	52.76	95.75	94.56	94.37
w/o GenLoss	78.28	76.32	75.12	62.98	57.50	54.34	96.17	94.85	94.90
GenFuse	**81.03**	**82.63**	**79.56**	**73.07**	**68.08**	**65.62**	**97.10**	**95.34**	**96.14**

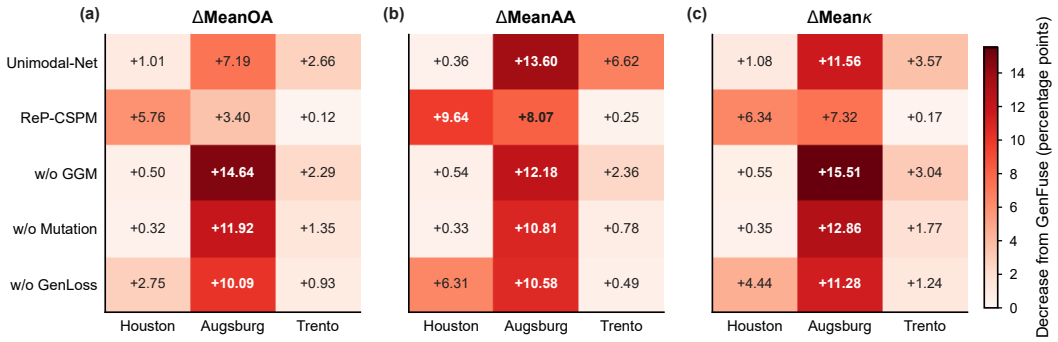


Figure 6: Complete structural-ablation visualization. Each cell reports the decrease in MeanOA, MeanAA, or Meanκ from full GenFuse under the same protocol as Table 6; larger values indicate a larger degradation.

CSPM modality parent. **w/o Mutation** retains feature-wise crossover but removes the range-scaled mutation offset. **w/o GenLoss** removes $\mathcal{L}_{\text{Gen}}$.

Table 6 preserves the absolute values, while Figure 6 makes the component-wise decreases explicit. The largest MeanOA loss occurs when GGM is removed on Augsburg (14.64 percentage points), followed by removing mutation (11.92 points) and genetic consistency loss (10.09 points). Replacing CSPM with a single nearest feature has its largest MeanOA effect on Houston (5.76 points). The Trento decreases are smaller, ranging from 0.12 to 2.66 points in MeanOA, consistent with the narrower margins among methods on this high-accuracy dataset. Because all entries average the same two deterministic missing directions, the heatmaps compare component effects rather than training-seed variability.

B.2 Alternative Proxy Strategies

The five classification variants isolate simpler ways to fill a missing representation. **Retrieval-only** directly uses the attention-weighted top-k modality parent as the proxy. **NN-only** uses the highest-scoring retrieved feature. **Generator-only** removes bank retrieval and combines the semantic parent with a learned proxy vector for the missing modality. **Mean imputation** uses the running mean of observed training features from the missing modality. The full **GenFuse** combines cross-sample retrieval, feature-wise crossover, and range-scaled mutation.

The directional view reveals why MeanOA alone is insufficient for interpreting the proxy strategies. On Augsburg, Retrieval-only reaches 89.40% OA when SAR is missing but falls to 36.40% when HSI is missing. GenFuse raises the latter direction to 59.13% and obtains the highest MeanOA of 73.07%, although Retrieval-only remains higher in the SAR-missing direction. On Trento, GenFuse is strongest in both directions, whereas Generator-only falls to 59.91% OA when HSI is missing despite reaching 98.50% when LiDAR is missing. Thus, none of the simpler strategies provides uniformly strong classification across datasets and missing directions.

Table 7: Direction-level alternative proxy-strategy results (%) under the specific modality-missing protocol with $\eta_{\text{train}} = 30\%$. MeanOA is the arithmetic mean of the two OA values, not a repeated-run statistic.

Dataset	Method	HSI missing OA/κ	LiDAR/SAR missing OA/κ	MeanOA
Augsburg	Retrieval-only	36.40/22.04	**89.40/85.23**	62.90
	NN-only	46.45/29.83	87.68/83.03	67.07
	Generator-only	53.78/40.61	85.64/80.15	69.71
	Mean imputation	6.48/−0.26	87.69/82.98	47.09
	GenFuse	**59.13/49.10**	87.00/82.13	**73.07**
Trento	Retrieval-only	92.63/90.28	98.40/97.87	95.52
	NN-only	93.89/91.93	98.45/97.93	96.17
	Generator-only	59.91/51.80	98.50/97.99	79.21
	Mean imputation	93.40/91.29	97.91/97.21	95.66
	GenFuse	**95.22/93.64**	**98.98/98.64**	**97.10**

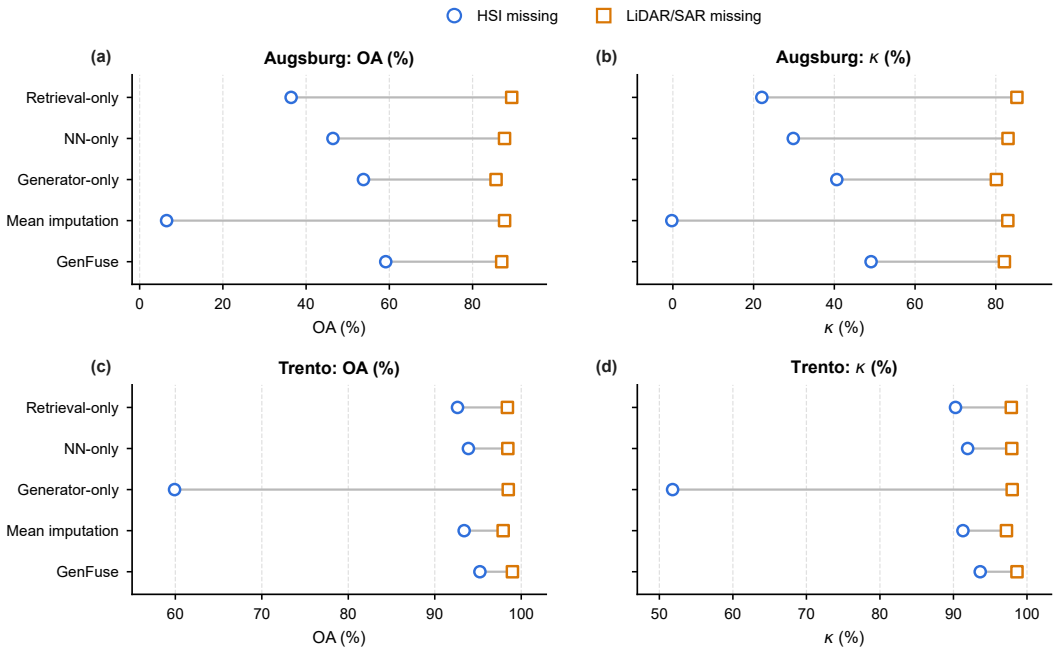


Figure 7: Direction-level comparison of alternative proxy strategies. Circles denote HSI-missing tests, squares denote LiDAR/SAR-missing tests, and each connector joins the two deterministic test directions of the same method. Exact OA and κ values are given in Table 7.

B.3 Robustness Checks on Houston

We repeat one challenging condition with three independent seeds: Houston under the specific modality-missing protocol, $\eta_{\text{train}} = 70\%$, and LiDAR removed at test time. The labeled split is fixed with `split_seed`=2, while initialization, data order, and training masks vary jointly with `run_seed` and `mask_seed` in $\{0, 1, 2\}$. For each seed, GenFuse and DPMamba use protocol manifests with identical hashes, confirming the same split coordinates and sample-wise training masks. All runs use the checkpoint from the fixed final training epoch; no run is selected by test performance.

To assess the effect of same-sample retrieval during complete-sample proxy alignment, we repeat GenFuse under the same three seeds while excluding every historical bank entry sharing the query sample ID. The original and exclusion branches use identical splits, missing masks, initialization,

data order, optimization schedules, and final-epoch checkpoint rules; only the candidate exclusion rule differs.

Table 8: Three-seed robustness checks (%) on Houston with 70% incomplete training samples and LiDAR removed at test time. Values are means ± sample standard deviations ($N = 3$, `ddof`=1).

Setting	OA	AA	κ
DPMamba	38.91 ± 18.38	47.54 ± 15.34	34.79 ± 18.72
GenFuse	92.13 ± 0.70	93.66 ± 0.63	91.49 ± 0.76
GenFuse w/ exclusion	92.96 ± 0.98	94.30 ± 0.80	92.38 ± 1.06

The paired OA changes after same-ID exclusion are $+0.05$, $+1.95$, and $+0.47$ percentage points for seeds 0, 1, and 2, respectively; no paired run decreases. Instrumentation confirms that same-ID candidates occur only for complete-sample proxy alignment: incomplete-sample queries select no same-ID entry in either stage. The exclusion branch selects no same-ID entry, and every query retains at least k eligible candidates. Because only three seeds are evaluated, the positive differences are reported descriptively and are not interpreted as statistically significant improvements.

C Retrieval Validity and Proxy-Feature Alignment

C.1 Retrieval-Validity Protocol

The retrieval-validity analysis in Figure 5(a–b) of the main paper is separate from CSPM's runtime FIFO banks. It re-encodes the complete training split into a fixed pool and evaluates 240 Trento training queries per direction. The query's same-index pool entry is excluded. Both runtime CSPM and this analysis rank candidates by the raw inner product without pre-retrieval ℓ_2 normalization. The learned query and key projections in CSPM are applied only after top-k selection to compute attention weights. Labels are consulted only after retrieval to calculate the same-class candidate ratio. Random retrieval samples candidates from the same fixed pool under the same exclusion rule.

Table 9 prevents the fixed-pool analysis from being interpreted as the runtime memory mechanism. The analysis pool, labels, and same-index exclusion are used only to measure retrieval validity. They neither alter CSPM candidate selection nor enter training or inference.

Figure 8 exposes the rate-specific values that are averaged in the main paper. For HSI queries against the LiDAR bank, inner-product retrieval ranges from 62.60–66.67%, 69.58–73.29%, and 93.54–97.92% across the 30%, 50%, and 70% models, respectively. For the reverse direction, the corresponding ranges are 63.33–65.87%, 81.67–90.16%, and 89.71–91.67%. Random retrieval remains between 13.75% and 19.17% over all panels. These results establish a consistent separation from the random reference, but comparisons between training missing rates remain descriptive because each column uses a different trained model.

C.2 Proxy-Feature Alignment

For each direction, a proxy is compared with the corresponding observed feature of the modality treated as missing in the analysis. This feature is used only for the alignment measurement. **Retrieval-only proxy** is the attention-weighted top-k aggregation, **Top-k mean proxy** is the unweighted candidate mean, **NN-only proxy** is the highest-scoring candidate, and **GenFuse proxy** is the output after crossover and range-scaled mutation. These four constructions are computed within one analysis pipeline and are distinct from the five independently trained classification variants in Section B.

GenFuse has the highest mean alignment (0.7239) and smallest query-level SD (0.0952) for HSI→LiDAR. The retrieval-based alternatives have broader distributions with SDs of 0.4219–0.4474. In the reverse direction, the Retrieval-only proxy has the highest mean cosine similarity (0.8756), whereas GenFuse has the smallest SD (0.1497) and attains the strongest direction-level OA in Table 7. Together, the table and distribution plot show that cosine similarity measures proxy-to-feature alignment rather than classification utility; a higher cosine value alone does not establish higher OA.

Table 9: Separation between runtime CSPM and the fixed-pool retrieval-validity analysis. Both use raw-inner-product ranking, but they differ in feature source, exclusion rules, label use, and update policy.

Property	Runtime CSPM	Retrieval-validity analysis
Feature source	FIFO banks populated from observed training features	Fixed pool obtained by re-encoding the complete training split
Matching	Raw inner product before attentive aggregation	Raw inner product
Same-index exclusion	Not applied as an additional runtime rule	Query's same-index pool entry is excluded
Label use	Labels do not enter retrieval	Labels are consulted only after retrieval to score same-class ratios
State update	Updated in Stage 1 and frozen afterward	Constructed once for the analysis and never updated
Use at inference	Frozen banks support proxy construction	Analysis pool is not used by the submitted classifier

Table 10: Same-class candidate ratios (%) on Trento for the two retrieval directions. Each rate uses the corresponding $\eta_{\text{train}} \in \{30\%, 50\%, 70\%\}$ model; the main-paper curves are arithmetic means of these rate-specific values.

Direction	k	30%		50%		70%	
		Inner product	Random	Inner product	Random	Inner product	Random
HSI→LiDAR	1	66.67	19.17	72.08	14.58	97.92	16.25
	4	62.60	15.94	69.58	15.83	95.42	18.44
	8	64.11	16.35	70.36	15.63	93.54	16.15
	16	64.82	16.74	71.72	16.80	95.42	16.85
	32	64.34	15.69	73.29	17.25	96.11	16.61
LiDAR→HSI	1	63.33	15.83	81.67	18.75	91.25	13.75
	4	63.54	18.44	83.96	17.29	91.67	16.77
	8	65.00	16.98	88.49	16.61	90.78	15.73
	16	65.05	15.26	90.16	17.01	90.99	16.35
	32	65.87	16.91	89.66	16.29	89.71	16.17

D Sensitivity and Efficiency

D.1 Training Missing-Rate Sensitivity

Figure 10 reports the observed response to $\eta_{\text{train}} \in \{0\%, 30\%, 50\%, 70\%, 90\%\}$. Houston decreases gradually from 67.40% to 63.82% without HSI and from 97.07% to 93.05% without LiDAR. Augsburg is strongly direction dependent: SAR-missing OA remains within 86.17–89.21%, whereas HSI-missing OA varies between 40.55% and 80.75% and is non-monotonic. Trento remains within 94.66–95.36% without HSI and 97.30–98.98% without LiDAR. These observations characterize the measured response curves and do not identify a causal regularization mechanism.

D.2 Candidate Set and Bank-Size Sensitivity

Table 12 reports the candidate-set sweep under the specific protocol with $\eta_{\text{train}} = 30\%$. The primary specific-protocol configuration uses $k = 32$ for HSI-missing tests and $k = 4$ for LiDAR/SAR-missing tests. The results vary by dataset and missing direction; no single candidate-set size is uniformly optimal.

The candidate-set sweep is direction dependent. Houston changes by at most 0.56 points without HSI, while $k = 8$ gives the highest LiDAR-missing OA (96.23%). Augsburg without HSI improves from 39.10% at $k = 4$ to 68.00% at $k = 16$ before decreasing to 59.13% at $k = 32$; the SAR-missing values remain within 86.42–87.31%. Trento without HSI increases to 95.22% at $k = 32$, whereas its LiDAR-missing maximum occurs at $k = 4$. The reported experiments apply the same

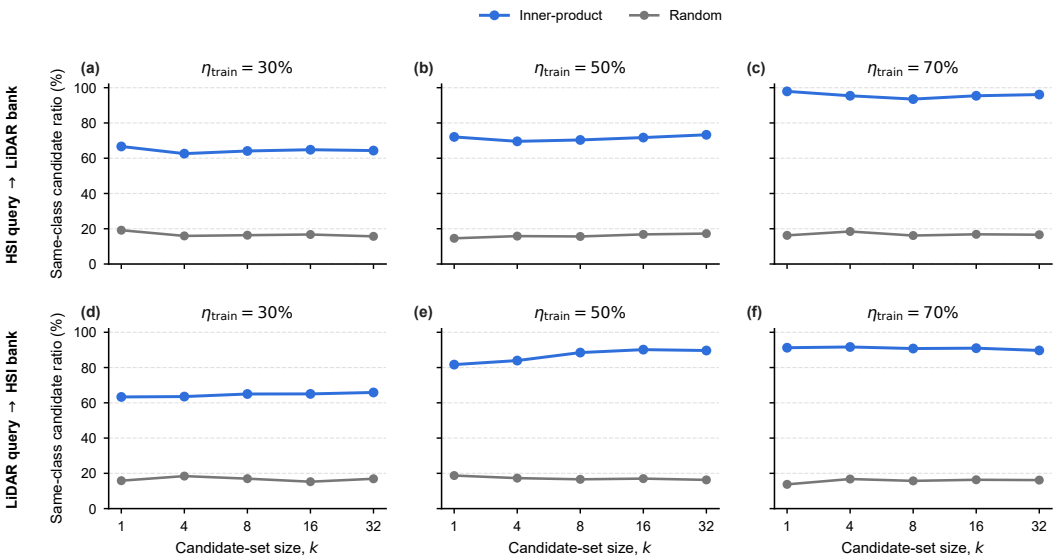


Figure 8: Rate-specific retrieval-validity results on Trento. Rows correspond to the two query-to-bank directions and columns to models trained with 30%, 50%, and 70% missing samples. Blue and gray curves report inner-product and random retrieval, respectively; no averaging or error bars are applied within these panels.

Table 11: Sample-level cosine similarity between each proxy and the corresponding observed feature on Trento. SD and quantiles summarize $n = 240$ fixed-pool queries per direction; they are not training-seed statistics.

Direction	Proxy	Mean	SD	Median	Q1–Q3	Min.	Max.
HSI→LiDAR	Retrieval-only proxy	0.6782	0.4219	0.9635	0.2863–0.9977	−0.1962	0.9998
	Top-k mean proxy	0.6911	0.4227	0.9733	0.3898–0.9971	−0.2299	0.9998
	NN-only proxy	0.6573	0.4474	0.9397	0.3483–0.9841	−0.3296	0.9997
	GenFuse proxy	**0.7239**	**0.0952**	0.7171	0.6518–0.8079	0.4268	0.8580
LiDAR→HSI	Retrieval-only proxy	**0.8756**	0.2534	**0.9942**	0.9625–0.9980	−0.3567	0.9999
	Top-k mean proxy	0.7619	0.3296	0.9901	0.5065–0.9981	−0.3776	0.9999
	NN-only proxy	0.7744	0.3048	0.9690	0.5415–0.9899	−0.3132	0.9999
	GenFuse proxy	0.6868	**0.1497**	0.7140	0.6270–0.7812	−0.0034	0.8643

direction-level rule across datasets ($k = 32$ for HSI missing and $k = 4$ for LiDAR/SAR missing) instead of selecting the best value separately for each dataset from this sweep.

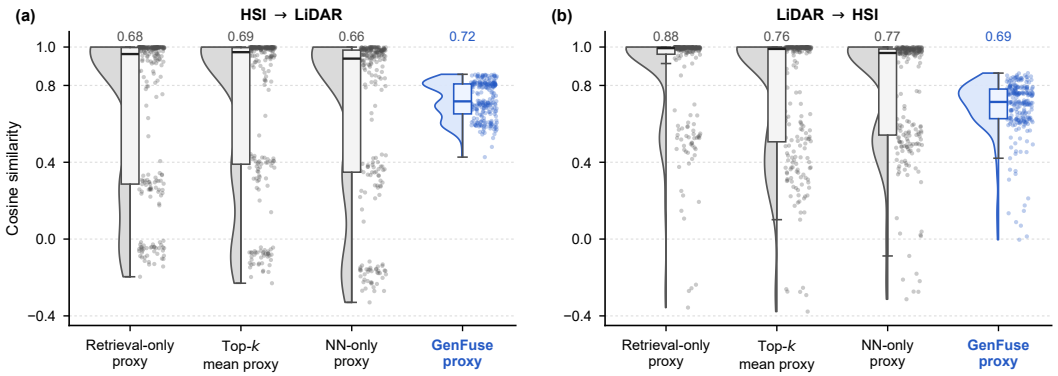


Figure 9: Sample-level proxy alignment on Trento. Half violins, box plots, and jittered points show the 240 observed cosine values per proxy and direction; numbers above the distributions are arithmetic means. The distributions are drawn from the measured samples rather than synthesized from summary statistics.

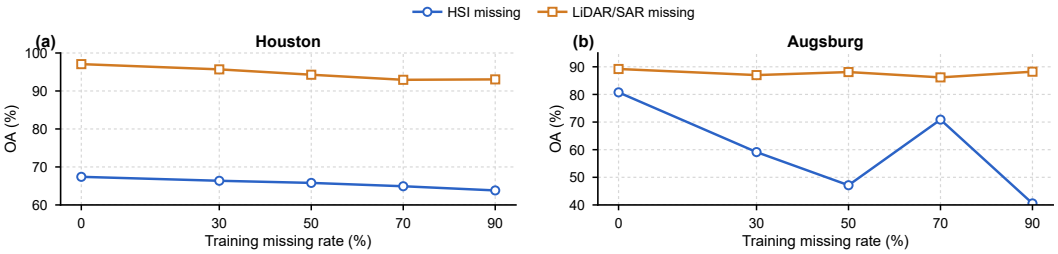


Figure 10: OA versus η_{train} under the specific modality-missing protocol. Each point uses a deterministic missing modality at test time. Panels use dataset-specific vertical ranges to resolve within-dataset variation.

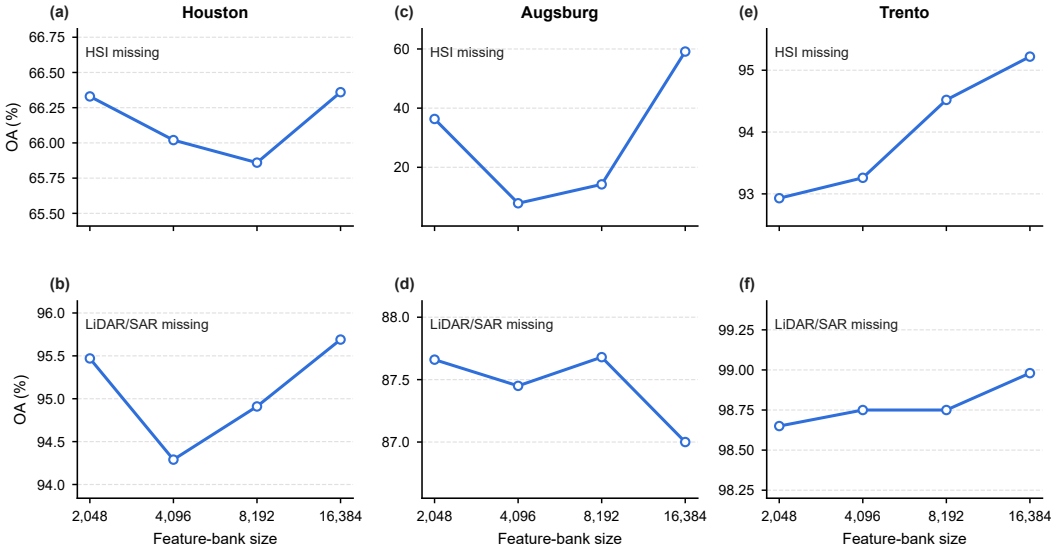


Figure 12: OA sensitivity to feature-bank size under the same protocol as Table 13. Columns correspond to datasets and rows to deterministic missing directions. Panels use dataset- and direction-specific vertical ranges, while the table provides the complete OA/AA/κ values.

Table 12: Candidate-set sensitivity at $\eta_{\text{train}} = 30\%$ under the specific modality-missing protocol. Each entry is OA/AA/κ (%).

Dataset	Missing modality	$k = 4$	$k = 8$	$k = 16$	$k = 32$
Houston	HSI	66.19/68.32/63.59	66.08/68.41/63.48	65.80/67.73/63.15	66.36/68.73/63.77
Houston	LiDAR	95.69/96.52/95.34	96.23/96.94/95.92	95.08/95.95/94.68	95.72/96.50/95.37
Augsburg	HSI	39.10/30.17/25.63	60.53/46.52/48.54	68.00/56.00/58.46	59.13/56.21/49.10
Augsburg	SAR	87.00/79.95/82.13	86.42/79.77/81.38	86.65/79.54/81.68	87.31/80.03/82.52
Trento	HSI	92.93/92.35/90.66	93.39/92.64/93.39	94.13/92.91/92.23	95.22/92.82/93.64
Trento	LiDAR	98.98/97.85/98.64	98.85/98.24/98.47	98.00/98.09/98.39	98.78/98.06/98.36

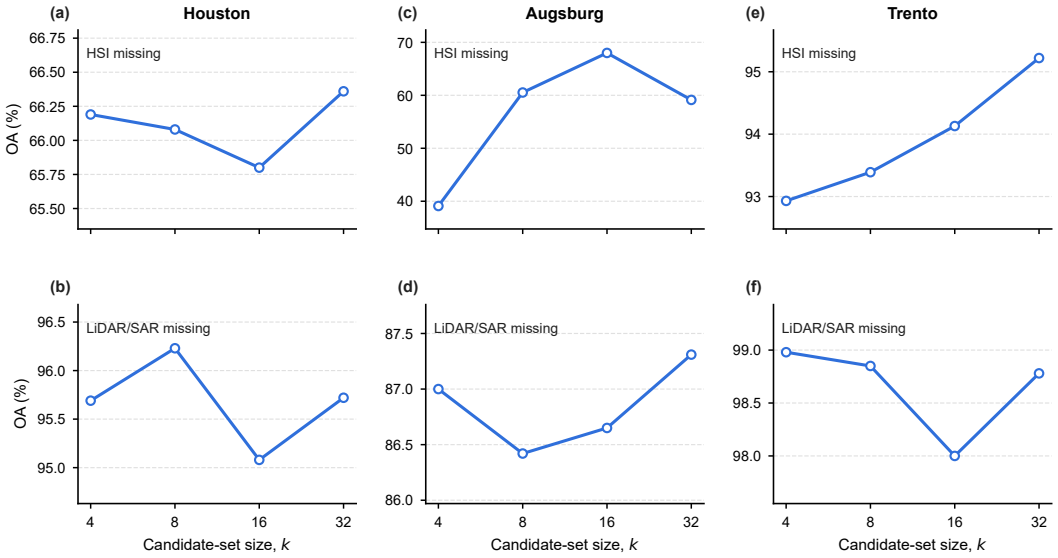


Figure 11: OA sensitivity to candidate-set size under the same protocol as Table 12. Columns correspond to datasets and rows to deterministic missing directions. Dataset- and direction-specific vertical ranges expose within-panel variation; exact OA/AA/κ values are retained in the table.

The bank-size results are also direction dependent. Houston varies within 0.50 points without HSI and 1.40 points without LiDAR. Augsburg without HSI is strongly non-monotonic, attaining 36.34%, 7.84%, 14.23%, and 59.13% OA as capacity increases; its SAR-missing results remain within 87.00–87.68%. Trento without HSI increases from 92.93% to 95.22%, whereas the LiDAR-missing values differ by only 0.33 points. We report these measured patterns directly rather than assigning a causal explanation without additional controlled evidence. All primary experiments retain the fixed capacity of 16,384.

D.3 Measured Efficiency

GenFuse requires 201.984 ms per training batch, compared with 77.118 ms without retrieval and 117.694 ms for NN-only. Missing-input inference remains below 0.17 ms per sample for all variants: 0.1618 ms for GenFuse, 0.1231 ms without retrieval, and 0.1321 ms for NN-only. The two 16,384-entry banks occupy 8.0 MiB, while peak allocated VRAM differs by 8.69 MiB between GenFuse and the retrieval-free variant and by 0.72 MiB between GenFuse and NN-only. These measurements quantify the added retrieval and proxy-generation cost without treating a single measured setup as a repeated-run estimate. For timing, the original two-batch training loader was cycled within the measurement procedure, whereas the inference loader was not cycled.

Table 13: Feature-bank-size sensitivity at $\eta_{\text{train}} = 30\%$ under the specific modality-missing protocol. Each entry is OA/AA/κ (%).

Dataset	Missing modality	2,048	4,096	8,192	16,384
Houston	HSI	66.33/68.52/63.75	66.02/67.92/63.40	65.86/67.85/63.23	66.36/68.73/63.77
Houston	LiDAR	95.47/96.23/95.10	94.29/95.31/93.83	94.91/95.88/94.49	95.69/96.52/95.34
Augsburg	HSI	36.34/32.07/25.70	7.84/16.21/3.95	14.23/32.33/9.92	59.13/56.21/49.10
Augsburg	SAR	87.66/78.68/82.92	87.45/78.25/82.64	87.68/78.48/82.94	87.00/79.95/82.13
Trento	HSI	92.93/92.35/90.66	93.26/92.56/91.10	94.52/93.57/92.73	95.22/92.82/93.64
Trento	LiDAR	98.65/97.87/98.20	98.75/97.94/98.33	98.75/98.02/98.33	98.98/97.85/98.64

Table 14: Measured efficiency on Trento; nonzero banks store `float32` features with $d = 64$. Measurements use Stage 1, PCA preprocessing, input patch size 15, $\eta_{\text{train}} = 30\%$, 40 training samples per class, seed 2, batch size 120, bank size 16,384, and $k = 32$. Training time includes forward, loss, backward, and optimizer steps. Inference includes `eval()`, `no_grad()`, forward, softmax, and argmax. Five warm-up and 30 timed batches were used per variant and inference mode; missing-input time is the arithmetic mean over the two missing directions. This is one measured setup, not independent repeated runs.

Method	Bank memory (MiB)	Train time (ms/batch)	Train throughput (samples/s)	Complete inference (ms/sample)	Missing inference avg. (ms/sample)	Peak VRAM (MiB)
GenFuse	8.0	201.984	594.11	0.3050	0.1618	2448.13
w/o retrieval	0.0	77.118	1556.06	0.2343	0.1231	2439.44
NN-only	8.0	117.694	1019.59	0.2493	0.1321	2447.41

E Qualitative Results and Evaluation Scope

E.1 Additional Classification Maps

In the highlighted Trento region, several baselines introduce scattered class changes inside otherwise contiguous land-cover areas. GenFuse produces a more coherent prediction pattern with fewer isolated pixels, consistent with its high direction-level OA. This visual comparison is qualitative and is interpreted together with the quantitative results rather than as an independent accuracy measure.

The highlighted Houston region contains thin, spatially localized classes that are easily fragmented in the classification maps. GenFuse preserves a more continuous highway pattern and reduces scattered substitutions relative to the compared methods. As with the Trento map, this observation illustrates spatial prediction behavior under one fixed test condition and does not replace the reported OA, AA, or κ.

E.2 Scope of the Evaluation

The reported experiments use sample-wise zero masking of otherwise aligned bimodal inputs. They do not establish performance under real sensor outages, spatial misalignment, temporal inconsistency, domain shift, detection or segmentation tasks, or three-modality fusion. These settings require separate data and controlled protocols. The present results therefore support the stated incomplete multimodal training setting for bimodal remote-sensing classification.

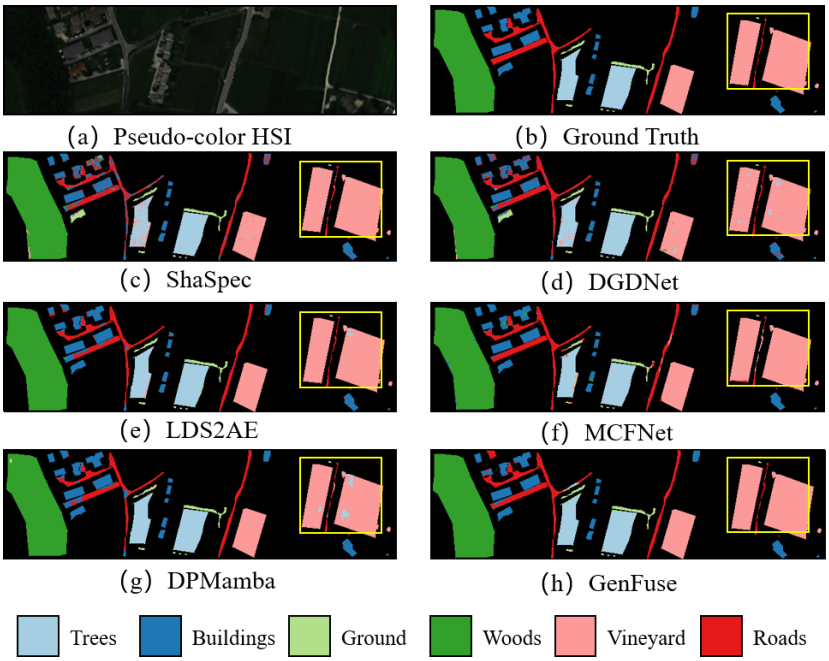


Figure 13: Classification maps on Trento under the specific modality-missing protocol with $\eta_{\text{train}} = 70\%$ and LiDAR removed at test time. Highlighted regions support local comparison of spatial prediction patterns.

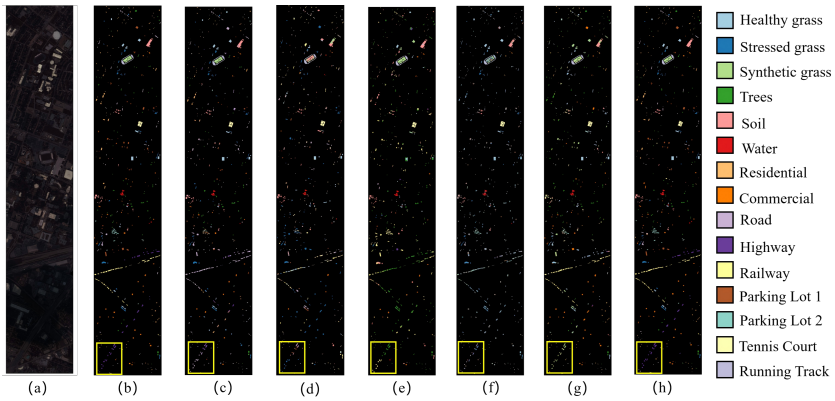


Figure 14: Classification maps on Houston under the specific modality-missing protocol with $\eta_{\text{train}} = 30\%$ and LiDAR removed at test time. Highlighted regions support local comparison among methods.

F Encoder and Projection Architecture

Table 15 gives the implementation-level dimensions of each modality encoder. Let P denote the spatial width of the input patch listed in Tables 3 and 4, so the token count is $N = P^2$ because the embedding kernel and stride are both 1×1. The HSI and LiDAR/SAR encoders have the same topology but do not share parameters.

Table 15: Implementation-level architecture of each modality encoder. P is the dataset-specific input patch width, $N = P^2$, and B is the batch size. The two modality encoders use the same architecture without sharing weights.

Stage	Layer	Configuration	Output shape
Input	Modality patch	C_{in} channels, spatial size $P \times P$	(B, C_{in}, P, P)
Spectral reduction	Conv block 1	1×1 Conv $(C_{\text{in}} \to 128)$, BN, ReLU	$(B, 128, P, P)$
	Conv block 2	1×1 Conv $(128 \to 128)$, BN, ReLU	$(B, 128, P, P)$
Token embedding	Linear projection	1×1 Conv $(128 \to 64)$, stride 1	$(B, N, 64)$
Token sequence	Position and class tokens	Trainable class token; 2D sine–cosine initialization	$(B, N+1, 64)$
Encoder	Transformer blocks	4 blocks, 8 heads, MLP ratio 2.0, QKV bias	$(B, N+1, 64)$
Global summary	Class-token extraction	LayerNorm followed by index-0 token	$(B, 64)$
Projection head	Nonlinear MLP	Linear$(64, 64)$–SiLU–Linear$(64, 64)$	$(B, 64)$

Patch-embedding and linear-layer weights use Xavier-uniform initialization, class tokens use a zero-mean normal initialization with standard deviation 0.02, and LayerNorm scale and bias are initialized to one and zero, respectively. The positional embeddings are initialized from a two-dimensional sine–cosine pattern and remain trainable. The modality-specific projection heads produce 64-dimensional task features used by CSPM, GGM, and the feature banks; their parameters are not shared.