
Harnessing Multi-Modal Co-learning for Missing Earth Observation Modalities

Francisco Mena

GFZ Helmholtz

Centre for Geosciences

Potsdam, Germany

Dino Ienco, Cassio F. Dantas

INRAE, UMR TETIS, Inria

University of Montpellier

Montpellier, France

Roberto Interdonato

CIRAD, UMR TETIS, Inria

University of Montpellier

Montpellier, France

Abstract

The increasing availability of multi-modal satellite data has advanced Earth observation (EO) research by enabling the analysis of complex environmental phenomena. However, the lack of systematically available sensor modalities at inference time—due to operational constraints, weather conditions, or sensor failures—poses a major challenge for multi-modal deep learning models. This work explores enhancing model robustness to such missing EO modality scenarios through multi-modal co-learning, where modality-dedicated models collaborate and share knowledge during training. Preliminary experiments on crop recognition tasks demonstrate that combining co-learning at the feature level with adaptive strategies at the decision level improves robustness under incomplete EO sensor availability.

1 Introduction

Nowadays, more than 9000 satellites orbit and collect data about our planet [1]. These remote observations, regarded as multi-modal data, have been vital to analyzing and studying complex phenomena on Earth [2]. Thus, advances in sensor technology and the accessibility of their collected data have increased the research on deep learning models using multimodal sensor data for Earth Observation (EO) applications [3, 4]. However, accessing multiple sensor modalities, also known as EO modalities, persistently and synchronously during both the training and inference stages could be infeasible in scenarios characterized by operational constraints.

The lack of systematically available sensor modalities covering the same region and period is an inherent problem in the EO domain [5]. This is because data collection occurs under operational constraints in real-world environments, affected by factors like limited spatial coverage, weather conditions, and unexpected errors. For instance, the Landsat 7 satellite experienced problems after 2003 (ETM+ SLC-off) [6], the NAIP satellite, which operates only in the US, and the Sentinel-1b satellite, which ceased operations in 2021 [7]. These situations affect the availability of EO modalities at inference and deployment time, named the missing modality problem.

In the deep learning field, missing modalities have been shown to greatly affect model performance [8], since even recent models are not naturally robust to missing data [9]. This scenario is illustrated in Fig. 1. In this way, the decline in performance is driven by the model’s robustness to missing data. For instance, Garnot et al. [10] observe that different fusion strategies make multi-modal models more robust to missing the optical modality. Moreover, Mena et al. [11] notice that cross-modal distillation improves model robustness compared to previous approaches in the literature. Thus, robust approaches and collaboration between modalities at the training stage are crucial for the missing modality problem in the EO domain.

Various multi-sensor models have been proposed in the EO literature to handle the missing modality problem. These models can be classified according to the which modalities are expected to be missing

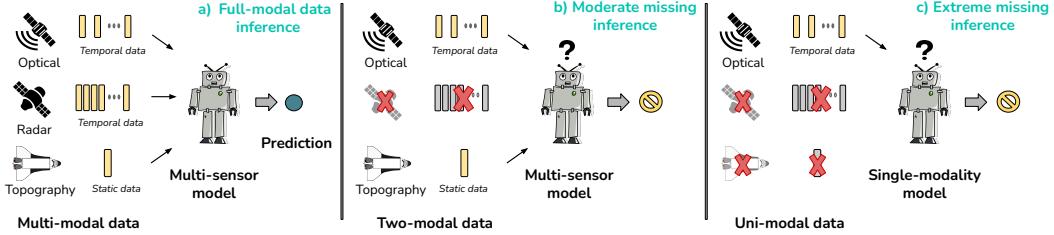


Figure 1: Illustration of the missing modality problem at inference time with EO data.

(or available) during inference [8]. For instance, the DisOptNet [12] model is pre-trained on optical images and then transferred to radar images with a hallucination branch. Similarly, Kampffmeyer et al. [13] use a hallucination branch from the optical image to imitate the expected missing depth image. On the other hand, some works consider simply sharing weight layers [14], or employing the Modality Dropout (ModDrop) technique [15, 11] to handle all kinds of missing modality (or sensor) cases. Besides, sensor-invariant modeling has been explored in Geospatial Foundation Models (GeoFM), such as AnySat [16] and Copernicus-FM [17]. Another classification angle is obtained by whether the approach recovers the lost modality. For instance, reconstruction-based GeoFMs [18, 19, 20] rely on a two-step process. First, recover the missing modalities (usually mono-temporal ones) by a cross-reconstruction process, and then perform the prediction.

Furthermore, the related literature evidence that methods effectively handling moderate missingness do not necessarily succeed in extreme missing conditions, and viceversa [11, 21]. This motivates our work to further enhance the robustness of multi-modal models to missing modalities. Our exploration relies on the multi-modal co-learning strategy [22, 23]. Thus, the objective is to have multiple models that share knowledge and cooperate to improve their individual learning and predictive capacity. In this way, our method does not recover the missing modalities; Instead, it performs direct inference with any set of modalities available. In contrast to previous reconstruction-based GeoFM, our method predicts without hallucinating inputs, avoiding error propagation and potentially distribution shift. Our preliminary results on crop recognition applications highlight the potential of our approach compared to recent literature for addressing the missing modality problem in the EO domain.

2 Method

Problem definition Let us consider the set of training modalities as \mathbb{M} , and the multi-modal input data as $\mathbb{X} = \{\mathbf{X}_v\}_{v \in \mathbb{M}}$. At inference time, any arbitrary subset of these modalities may be accessible $\tilde{\mathbb{M}} \subseteq \mathbb{M}$, expressed by $\tilde{\mathbb{X}} = \{\mathbf{X}_m\}_{m \in \tilde{\mathbb{M}}}$. This missing modality problem is illustrated in Fig. 1. To address this problem, we introduce a multi-modal model employing decision-level fusion [3]. Thus, unimodal-based predictions are averaged to yield the fused estimation, given by $\hat{y}_{\text{full}} = M^{-1} \cdot \sum_{m=1}^M \hat{y}_m$, and missing modalities are ignored from the merging when occurring. In addition, we employ a co-learning strategy driven by various loss functions. First, we use the standard cross-entropy as the main loss function $\mathcal{L}_{\text{main}}$ that guides the fused prediction (i.e., the multi-modal prediction). We also introduce loss functions at the feature- and decision-level as follows.

2.1 Feature-level co-learning

We assume that EO modalities have shared (invariant to modality) and specific (unique to each modality) information between them. For instance, a high-resolution and a low-resolution optical image have shared data (the optical part) and specific data (related to the differences in spatial resolutions). To model this, we use modality-dedicated encoders to extract both the specific and shared features explicitly per modality, expressed by $\mathbf{z}_m^{\text{sha}}, \mathbf{z}_m^{\text{sh}} = \mathcal{E}_m(\mathbf{X}_m)$.

For learning the specific features, we use a modality discriminant loss (with cross entropy), \mathcal{L}_{mod} . We consider a linear classifier that is fed with the specific features from either modality and has to predict the correct one. In this way, the specific features have to be distinguished between modalities.

For learning the shared features, we use a standard (cosine-based) contrastive loss [24], $\mathcal{L}_{\text{cont}}$. This is applied over the shared features of all pair-wise modalities, enforcing features to be similar across

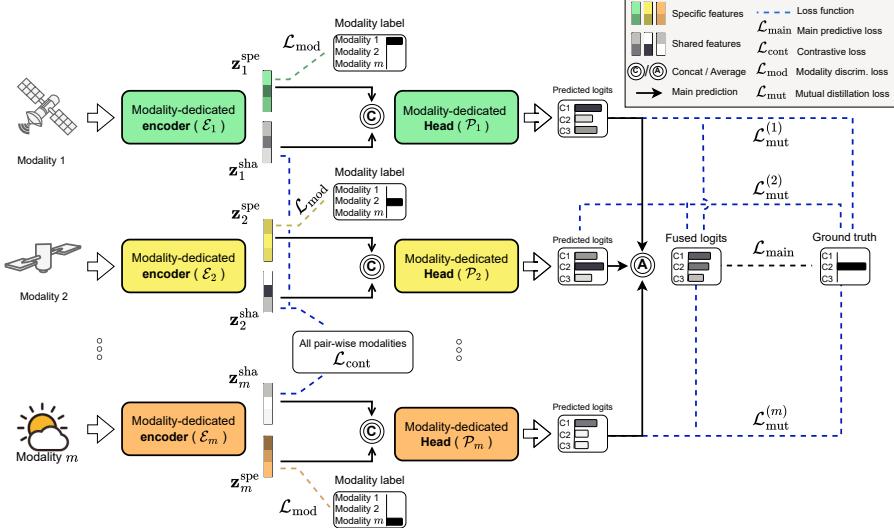


Figure 2: Illustration of the Full Co-learning (FullCo) method having four loss functions.

modalities. Thus, the positive pairs are the two modalities from the same sample, while the negative ones are the complementary shared features coming from all other samples in the batch.

To obtain the per-modality prediction, our model concatenates the specific and shared features and uses a modality-dedicated linear head. This is given by $\hat{y}_m = \mathcal{P}_m(\mathbf{z}_m^{\text{sha}} || \mathbf{z}_m^{\text{sha}})$.

2.2 Decision-level criteria

We consider two variants to handle the decision-level learning of the method.

Full Co-learning In this method, we follow a full co-learning strategy, where collaboration is enforced at both the feature and decision levels. The feature-level co-learning is discussed in Sec. 2.1. For the decision-level, we enforce that the modality-dedicated predictions \hat{y}_m go into the multi-modal consensus. To achieve this, we consider the mutual distillation approach introduced in [11], where the individual prediction per modality has to imitate the ground truth y as well as the fused prediction (consensus) \hat{y}_{full} . Thus, it can be seen as an extension of the DSensD+ method [11] by applying the co-learning at the feature-level as well and removing the sensor dropout. This method, illustrated in Fig. 2, is learned by optimizing an unweighted sum of all loss terms, given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{mod}} + \mathcal{L}_{\text{cont}} + \sum_{m=1}^M \mathcal{L}_{\text{mut}}^{(m)}. \quad (1)$$

Co-learning for Missing In this method, we also consider the feature-level co-learning introduced in Sec. 2.1. In addition, we incorporate the ModDrop at the decision-level to expose the model to a random subset of modalities during training. This is expressed by $\hat{y}_{\text{miss}} = M^{-1} \cdot \sum_{m \in \mathbb{M}} (1 - d_m) \cdot \hat{y}_m$, with $d_m \sim \text{Bern}(\alpha)$ the randomly drawn decision if modality m is masked out or not. Moreover, we enforce this prediction with missing modalities \hat{y}_{miss} to imitate the full-modal predictions \hat{y}_{full} , as well as the ground truth y , by a loss function $\mathcal{L}_{\text{miss}}$ named **missing distillation**. This method is illustrated in Fig. 3. Similar to FullCo, this method is learned by optimizing the following loss function

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \mathcal{L}_{\text{mod}} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{miss}}. \quad (2)$$

3 Experiments

Dataset For validation, we consider the crop recognition problem by using the CropHarvest benchmark [25]. Each sample has three temporal modalities at $10[m]$ resolution: multi-spectral optical, radar, and weather data. Besides, the samples have one mono-temporal modality, the topographic information. Since there is no test partition in this dataset, we use a standard 10-fold cross-validation and measure predictive performance by the weighted F1 (F1) score.

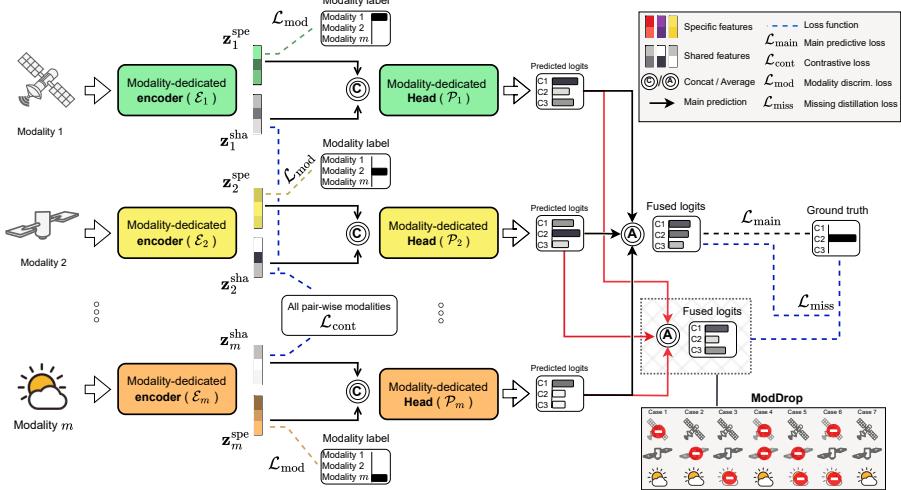


Figure 3: Illustration of the Co-learning for Missing (Co-Miss) method with four loss functions.

Table 1: Multi-class F1 score in the crop-type recognition task.* Averaged over available cases.

opt.	rad.	wea.	top.	Uni-modal	Anysat	Galileo	FCoM-av	TIMML	DSensD+	FullCo	Co-Miss
✓	✓	✓	✓		71.2	69.7	76.7	72.4	76.5	76.1	78.0
○	✓	✓	✓				66.6	63.1	65.9	66.1	66.9
✓	○	✓	✓				74.6	71.0	75.4	75.1	76.9
✓	✓	○	✓				76.2	72.7	76.4	76.3	77.7
✓	✓	✓	○				76.5	72.5	76.8	76.3	78.0
Moderate avg.							73.5	70.0	73.6	73.4	74.9
✓	○	○	○	71.0	70.9	70.5	73.8	71.5	75.3	76.0	76.3
○	✓	○	○	56.0	51.5	50.4	53.6	53.6	57.3	59.2	57.9
○	○	✓	○	46.7		48.3	42.8	44.3	48.4	50.2	48.8
○	○	○	✓	28.0		30.1	19.1	2.6	31.5	31.7	30.7
Extreme avg.				50.4	61.2*	49.8*	47.3	43.0	53.1	54.3	53.4
Overall avg.				50.4*	64.6*	53.8*	62.2	58.2	64.9	65.0	65.4

Results In Table 1 we display the F1 score of our methods compared against FCoM-av [21], TIMML [15], and DSensD+ [11]. We also include models specifically trained for the unimodal (missing) data: two fine-tuned GeoFM, Anysat [26] and Galileo [27], and single-modality baselines. We consider the moderate and extreme averages for each missing condition, respectively (see Fig. 1).

Analysis We notice various robustness advantages of our methods to missing EO modalities compared to state-of-the-art approaches. The FullCo method is more effective in extreme missing conditions, as expected, due to the loss functions applied in the individual per-modality predictions. On the other hand, the Co-Miss method is more effective in the moderate missing conditions, and in the overall score metrics (all cases averaged). This highlights the potential of the missing distillation strategy to adapt multi-modal models into arbitrary missing modality conditions for prediction.

4 Conclusion

The possible lack of sensor modalities at inference time limits the applicability of multi-modal models in the EO domain. In this preliminary work, we show that multi-modal co-learning at feature-level, combined with adaptive strategies at decision-level can enhance model robustness to missing sensor modalities. Instead of recovering the missing modalities, our approaches infer with any arbitrary modalities available. Future work will focus on extending the validation to other datasets, as well as providing a deep study on the advantages and limitations of both methods (like ablation studies).

Acknowledgments and Disclosure of Funding

F. Mena work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 414984028 – SFB 1404 FONDA.

References

- [1] Active satellite tle data and information. <https://orbit.ing-now.com/>. Accessed: 2025-10-15.
- [2] Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein. *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*. John Wiley & Sons, New York, 2021.
- [3] Francisco Mena, Diego Arenas, Marlon Nuske, and Andreas Dengel. Common practices and taxonomy in deep multi-view fusion for remote sensing applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 4797 – 4818, 2024.
- [4] Sancho Salcedo-Sanz, Pedram Ghamisi, María Piles, Martin Werner, Lucas Cuadra, A Moreno-Martínez, Emma Izquierdo-Verdiguier, Jordi Muñoz-Marí, Amirhosein Mosavi, and Gustau Camps-Valls. Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, 63:256–272, 2020.
- [5] Huanfeng Shen, Xinghua Li, Qing Cheng, Chao Zeng, Gang Yang, Huifang Li, and Liangpei Zhang. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):61–85, 2015.
- [6] Brian L Markham, James C Storey, Darrel L Williams, and James R Irons. Landsat sensor performance: History and current status. *IEEE Transactions on Geoscience and Remote Sensing*, 42(12):2691–2694, 2004.
- [7] Pierre Potin, Olivier Colin, Muriel Pinheiro, Betlem Rosich, Alistair O’Connell, Thomas Ormston, Jean-Baptiste Gratadour, and Ramón Torres. Status and evolution of the Sentinel-1 mission. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 4707–4710. IEEE, 2022.
- [8] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.
- [9] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186. IEEE, 2022.
- [10] Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:294–305, 2022.
- [11] Francisco Mena, Dino Ienco, Dantas F. Cassio, Roberto Interdonato, and Andreas Dengel. Multi-sensor model for Earth observation robust to missing data via sensor dropout and mutual distillation. *IEEE Access*, 13:83930 – 83943, 2025.
- [12] Jian Kang, Zhirui Wang, Ruoxin Zhu, Junshi Xia, Xian Sun, Ruben Fernandez-Beltran, and Antonio Plaza. DisOptNet: Distilling semantic knowledge from optical images for weather-independent building segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 2022.
- [13] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Urban land cover classification with missing data modalities using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):1758–1768, 2018.
- [14] Francisco Mena, Diego Arenas, and Andreas Dengel. Increasing the robustness of model predictions to missing sensors in Earth observation. *arXiv preprint arXiv:2407.15512*, 2024.
- [15] Guozheng Xu, Xue Jiang, Yue Zhou, Jia Fu, Zicheng Huang, and Xingzhao Liu. Transformer-based incomplete multi-modal learning for land cover classification. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 7276–7281. IEEE, 2024.
- [16] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-supervised modality fusion for Earth observation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2025.

- [17] Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos I. Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiaoxiang Zhu. Towards a unified Copernicus foundation model for Earth vision. *International Conference in Computer Vision*, 2025.
- [18] Gabriel Tseng, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.
- [19] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pages 164–182. Springer Nature Switzerland, 2024.
- [20] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. Accepted at *International Conference in Computer Vision*, 2025.
- [21] Francisco Mena, Diego Arenas, and Andreas Dengel. Missing data as augmentation in the Earth observation domain: A multi-view learning approach. *Neurocomputing*, 638, 2025.
- [22] Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. Foundations of multimodal co-learning. *Information Fusion*, 64:188–193, 2020.
- [23] Nhi Kieu, Kien Nguyen, Abdullah Nazib, Tharindu Fernando, Clinton Fookes, and Sridha Sridharan. Multimodal co-learning meets remote sensing: Taxonomy, state of the art, and future works. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [24] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.
- [25] Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. CropHarvest: A global dataset for crop-type classification. *Proceedings of NIPS Datasets and Benchmarks Track*, 2021.
- [26] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An Earth observation model for any resolutions, scales, and modalities. In *Accepted at the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [27] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. *International Conference on Machine Learning*, 2025.
- [28] Cassio F Dantas, Raffaele Gaetano, Claudia Paris, and Dino Ienco. Reuse out-of-year data to enhance land cover mapping via feature disentanglement and contrastive learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:1681–1694, 2024.
- [29] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

A Supplementary Material

Here we detail the loss functions used in both of our methods (FullCo and Co-Miss). First, the main predictive loss is given by

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{CE}}(y, \hat{y}_{\text{full}}) , \quad (3)$$

where $\mathcal{L}_{\text{CE}}(p, q) = -\sum_k \mathbb{1}(p = k) \log q_k$ with $\mathbb{1}(\cdot)$ the indicator function, and y is the ground truth. Then, the modality discriminant loss, used to learn the specific features, is given by

$$\mathcal{L}_{\text{mod}} = \sum_{m=1}^M \mathcal{L}_{\text{CE}}(m, \mathcal{P}^{\text{spe}}(\mathbf{z}_m^{\text{spe}})) , \quad (4)$$

where $\mathcal{P}^{\text{spe}}(\cdot)$ is an auxiliary linear layer to estimate the modality from the specific features. The contrastive loss, used to learn shared features, is given by

$$\mathcal{L}_{\text{cont}} = \sum_{m=1}^M \sum_{j \neq m} \mathcal{L}_{\text{info}}(\mathbf{z}_m^{\text{sha}}, \mathbf{z}_j^{\text{sha}}; \gamma) , \quad (5)$$

where $\mathcal{L}_{\text{info}}(\cdot, \cdot)$ is the InfoNCE [24] criteria based on the cosine similarity, parametrized by the temperature γ (we use 0.07 following [28]). The mutual distillation per modality m is given by

$$\mathcal{L}_{\text{mut}}^{(m)} = \mathcal{L}_{\text{CE}}(y, \hat{y}_m) + \lambda \cdot \mathcal{L}_{\text{KD}}(\hat{y}_{\text{full}}, \hat{y}_m; \tau), \quad (6)$$

where λ a weighting factor (we use τ^2 following [29]), $\mathcal{L}_{\text{KD}}(\cdot, \cdot; \tau)$ is the knowledge distillation function parametrized by the temperature τ (we use 0.5), following [11]. Finally, the missing distillation factor is given by

$$\mathcal{L}_{\text{miss}} = \mathcal{L}_{\text{CE}}(y, \hat{y}_{\text{miss}}) + \lambda \cdot \mathcal{L}_{\text{KD}}(\hat{y}_{\text{full}}, \hat{y}_{\text{miss}}; \tau), \quad (7)$$

where $\lambda = \tau^2$ and $\tau = 0.5$.