Semantically Conditioned Prompts for Visual Recognition under Missing Modality Scenarios *Supplementary Material*

Vittorio Pipoli[1,2], Federico Bolelli[1], Sara Sarto[1], Marcella Cornia[1],
Lorenzo Baraldi[1], Costantino Grana[1], Rita Cucchiara[1], and Elisa Ficarra[1]

[1]University of Modena and Reggio Emilia, Italy
[2]University of Pisa, Italy

{name.surname}@unimore.it

Additional Details on Missing Modality Scenarios. In Fig. 5, we aim to provide a visual representation of the missing modality scenarios considered in our experiments to enhance the clarity of mathematical notations used in the main paper. The diagram is divided into three primary sectors: *input modality state*, *missing modality scenarios*, and *missing modality cases*, which describe the problem from the finest to the coarsest granularity.

The input modality state outlines the potential availability of each modality for each sample, indicating whether a modality may be present or absent. The missing modality scenarios describe the state of an individual input sample, which can be complete (if all modalities are present), or have missing text or missing image modalities.

The missing modality cases, by contrast, outline the whole experimental setting. The specific case for each experiment must be defined at the outset. Once a case is selected, each sample in the data loader is assigned to one of its admissible missing modality scenarios, with the probabilities for each scenario predetermined. For example, if the selected missing modality case is missing text, then during training and inference, the samples provided by the dataloader may either be complete or missing text, but cannot be missing images.

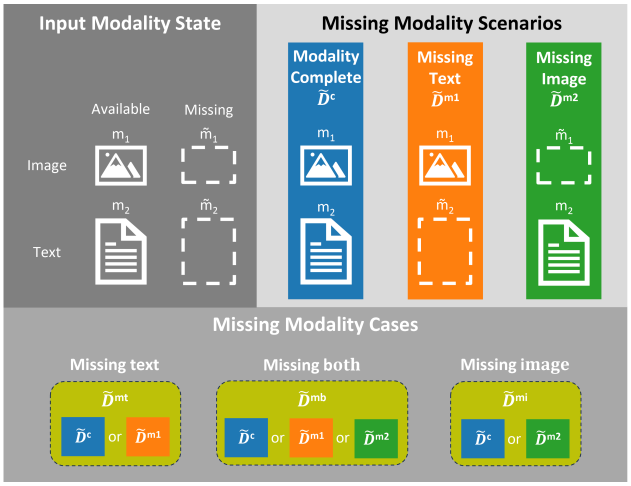
Input Modality State
Missing Modality Scenarios
Available
Missing
Image
Text
m_1
$\tilde{m}_1$
m_2
$\tilde{m}_2$
Modality Complete $\tilde{D}^c$
Missing Text $\tilde{D}^{m1}$
Missing Image $\tilde{D}^{m2}$
Missing Modality Cases
Missing text
Missing both
Missing image
$\tilde{D}^{mt}$
$\tilde{D}^{mb}$
$\tilde{D}^{mi}$
$\tilde{D}^c$ or $\tilde{D}^{m1}$
$\tilde{D}^c$ or $\tilde{D}^{m1}$ or $\tilde{D}^{m2}$
$\tilde{D}^c$ or $\tilde{D}^{m2}$

Figure 5. Visual diagram illustrating the potential input modality states (impacting each sample modality), missing modality scenarios (affecting each input sample), and missing modality case (impacting each experiment).

Table 4. AUC scores of the rightmost column of Fig. 6. TMMC stands for Train Missing Modality Case, which in this case can be missing-text or missing-image. The Train Missing Rate η is fixed at 70% for all the experiments.

Dataset	Metric	TMMC	Baseline	MAP	SCP
Food	AUC (Accuracy)	text image	69.50 77.65	$\underline{77.56}$ $\underline{87.22}$	**78.54** **87.49**
MM-IMDb	AUC (F1-Macro)	text image	36.80 39.93	$\underline{39.71}$ $\underline{46.89}$	**40.73** **49.31**
Hateful Memes	AUC (AUROC)	text image	60.99 $\underline{64.56}$	$\underline{61.14}$ 60.18	**61.43** **67.09**

Table 5. AUC scores of the rightmost column of Fig. 2 of the main paper. TMR stands for Train Missing Rate, which in this case can be 10%, 70%, or 90%. The Train Missing Modality Case is fixed to missing-both for the both train and test phases.

Dataset	Metric	TMR η	Baseline	MAP	SCP
Food	AUC (Accuracy)	10% 70% 90%	71.08 71.74 71.15	$\underline{80.01}$ $\underline{80.87}$ $\underline{81.07}$	**81.27** **81.57** **82.16**
MM-IMDb	AUC (F1-Macro)	10% 70% 90%	38.41 38.24 36.25	$\underline{42.40}$ $\underline{43.57}$ $\underline{39.72}$	**44.18** **44.80** **42.82**
Hateful Memes	AUC (AUROC)	10% 70% 90%	$\underline{62.93}$ $\underline{62.01}$ $\underline{62.10}$	61.54 60.69 52.21	**65.67** **64.00** **63.44**

Robustness to Different Missing Rates. We extend the experiment of *robustness to different missing rates* to the

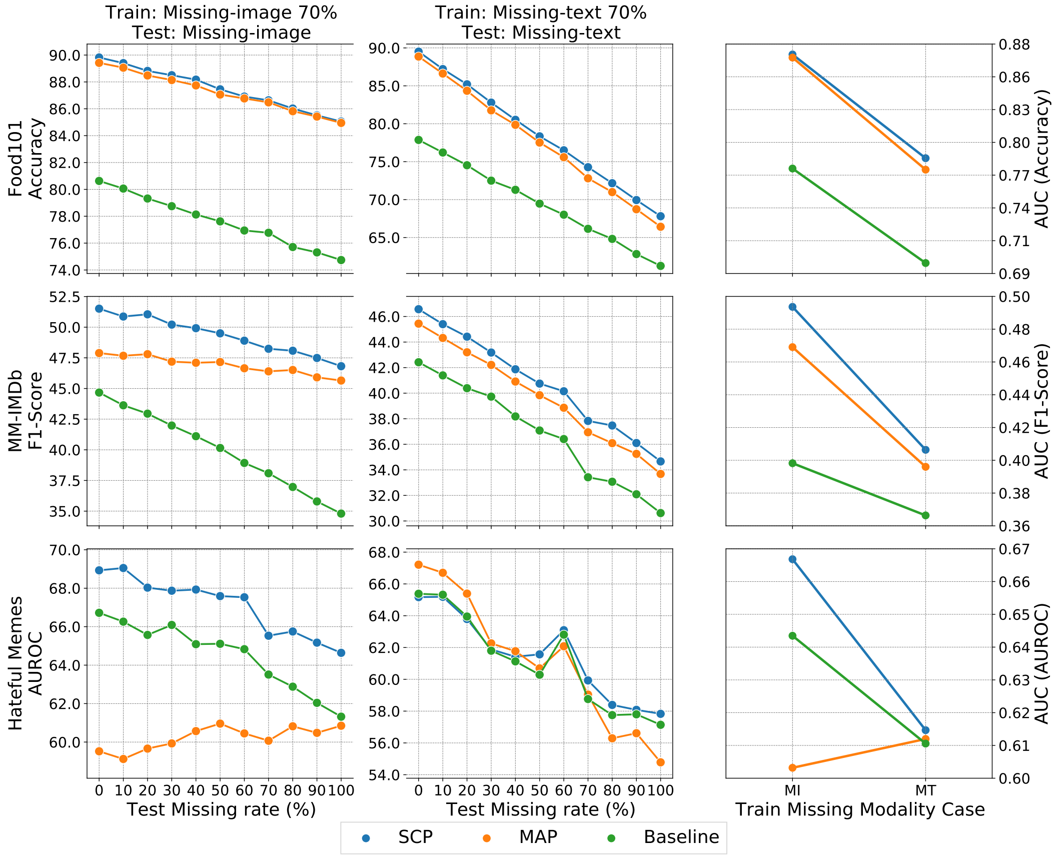

Figure 6. Robustness to different Train Missing Modality Cases and Test Missing Rates of SCP, MAP, and Baseline on Food101, MM-IMDb, and Hateful Memes. The Train Missing Rate is fixed at 70% for all the experiments.

other two missing modality cases, namely missing-text and missing-image. With that in aim, we evaluate the robustness of our proposal SCP with respect to our main competitor MAP [3] and Baseline. Specifically, we train the models with train missing rate 70% and then we test them at different missing rates varying them in a range from 0% to 100% with a step of 10% for both the missing-text and missing-image missing modality cases. The missing modality case in the testing phase is equal to the corresponding training phase for consistency. Results are presented in Fig. 6 for the Food101 [4], MM-IMDb [1], and the Hateful Memes [2] datasets. As the plots show, our SCP is the most robust model under all missing modality cases. Predictably, under the missing-text case, the performance of the models dropped significantly. This is to be expected as the text seems to be the dominant modality for these tasks. As reported in Fig. 3 of the main paper, to ensure a quantitative comparison of such curves, the subplots on the rightmost column depict the area under the curve of the respective performance curves on the left. A tabular version of the aforementioned AUC scores can be found in Tab. 4. The tabular version of the AUC scores of the results reported in the main paper (Fig. 3) are presented in Tab. 5.

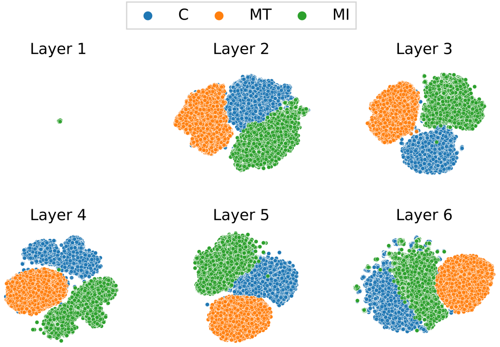

Figure 7. t-SNE visualization of the attention that the [CLS] and subsequent hidden states $\texttt{[CLS]}^i$ pay to the agnostic prompts for the first 6 layers of the ViLT architecture. Such ViLT architecture only harnesses agnostic prompts without SCP. In this way, the contribution of agnostic prompts is isolated from the semantically conditioned ones.

Visualization of Attention Patterns with t-SNE. We repeat the t-SNE experiment for agnostic prompts employing a model that only harnesses agnostic prompts without SCP. In this way, we further isolate the contribution of the agnostic prompts. With that said, we collect the attention weights corresponding to the attention that the

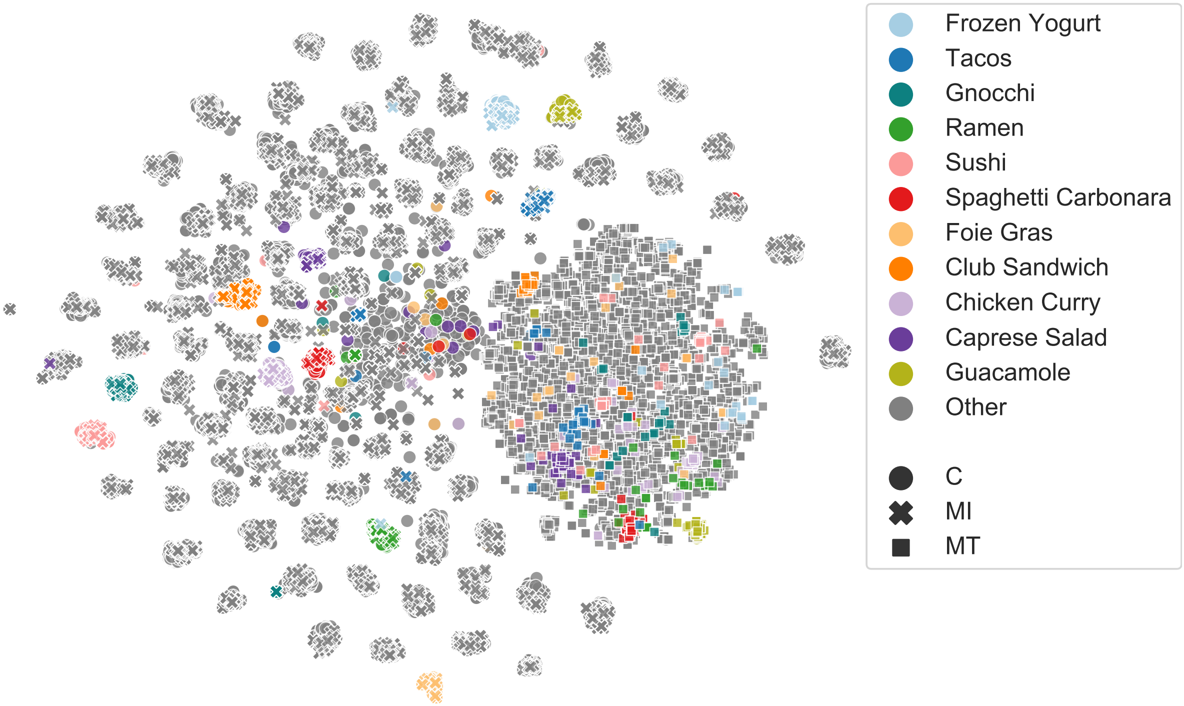

Figure 8. Attention weights of the semantically conditioned prompt generator module of SCP for the Food101 [4] test set. We provide only the annotation of the first 10 classes to reduce confusion in the plot.

`[CLS]` token and subsequent hidden states $\texttt{[CLS]}^i$ pay to the agnostic prompts across the first six layers of the ViLT architecture. The t-SNE visualization of such attention weights is presented in Fig. 7. The aforementioned figure showcases that a pool of agnostic prompts can automatically adjust itself to tackle different modality cases, without any manual adjustment. The chart shows that modality-complete (blue), missing-image (green), and missing-text (orange) lie in three different clusters, confirming that no external information about the missing modality scenario is required. Finally, it is impossible to spot patterns in the first layer because the `[CLS]`, before interacting with the available tokens, is the same for all the data samples, hence its attention patterns are always the same independently from the other tokens and or prompts, making the t-SNE representation collapse.

We offer an enhanced visualization of the SCP t-SNE analysis (Fig. 4c of the main paper) in Fig. 8. Notably, t-SNE is used to represent the attention weights of the semantically conditioned prompt generator of SCP of each test sample of Food101 [4]. As the chart shows, a big cluster corresponding to missing-text (squares, on the right) is clearly distinguishable from smaller clusters corresponding to modality-complete (circles) and missing-image (crosses), specialized on the input semantic. Within the big missing-text cluster, is it possible to spot some semantic subclusters, even if they are fuzzier with respect to their missing-image or modality-complete counterparts. We expect such a phenomenon because SCP relies on the efficacy of the $\texttt{[CLS]}^i$ representations to properly work. Indeed, the missing-text scenario leads to $\texttt{[CLS]}^i$ with weaker semantics, thus negatively affecting the attention mechanism of SCP.

For the sake of avoiding confusion both in the plot and in the legend, we provide a detailed annotation only for the first 10 classes of the dataset and we aggregate the remaining 90 classes in the dummy class *Other*.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated Multimodal Units for Information Fusion. In *Proceedings of the International Conference on Learning Representations Workshops*, 2017. 2

[2] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, 2020. 2

[3] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal Prompting with Missing Modalities for Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[4] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops*, 2015. 2, 3