Semantically Conditioned Prompts for Visual Recognition under Missing Modality Scenarios

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

Abstract

This paper tackles the domain of multimodal prompting for visual recognition, specifically when dealing with missing modalities through multimodal Transformers. It presents two main contributions: (i) we introduce a novel prompt learning module which is designed to produce sample-specific prompts and (ii) we show that modality-agnostic prompts can effectively adjust to diverse missing modality scenarios. Our model, termed SCP*, exploits the semantic representation of available modalities to query a learnable memory bank, which allows the generation of prompts based on the semantics of the input. Notably,* SCP *distinguishes itself from existing methodologies for its capacity of self-adjusting to both the missing modality scenario and the semantic context of the input, without prior knowledge about the specific missing modality and the number of modalities. Through extensive experiments, we show the effectiveness of the proposed prompt learning framework and demonstrate enhanced performance and robustness across a spectrum of missing modality cases. Our source code is available at* https://github.com/vittoriopipoli/SCP_WACV2025.

1. Introduction

Emulating human perceptual abilities has been a driving force of Deep Learning research. Over the years, the aspiration to mirror such a rich sensory integration has catalyzed the development of effective computational models capable of processing and relating information from diverse modalities [23, 26, 33]. Such models, referred to in the literature as *multimodal models*, are characterized by a *fusion step*, the critical point where the information from different modalities is combined to enable multimodal interactions. The fusion step determines the granularity at which the interactions between different modalities can be modelled by subsequent layers [20]. The literature distinguishes multimodal models mainly in *early fusion* ones, such as ViLT [14], and *late fusion* ones, like VATT [1], CLIP [27], and ImageBind [10]. The former models usually feed the concatenation of raw features coming from different modalities to a single encoder, which can thus harness fine-grained multimodal interactions. In contrast, the latter family feeds each modality to a specific encoder and then combines all the output feature vectors with a shallow network which models coarse-grained multimodal interactions [4].

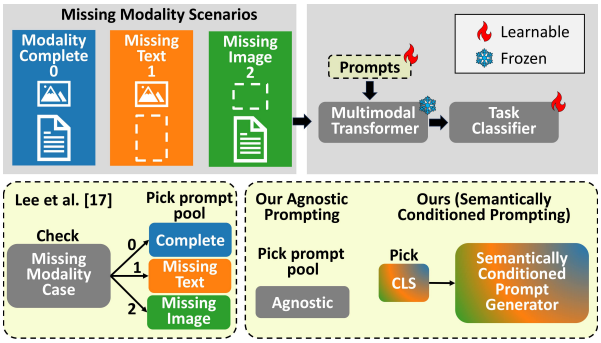


Figure 1. Top, from left to right: illustration of different missing modality scenarios: *modality-complete* (0), *missing-text* (1), and *missing-image* (2); the architecture employed as a basis for our contributions, ViLT [14]. Bottom: comparison between prompt-learning strategies for missing modalities – Lee *et al.* [16] and our SCP module, which generates sample-specific prompts.

Despite the recent advances, multimodal learning also comes with its own challenges [20]. The absence of one or more expected data streams (known as *missing modality*) is one of the most pressing issues and represents a critical barrier to the deployment of robust multimodal systems. The assumption of data completeness, which is usually made during training, is indeed frequently violated in real-world scenarios where models encounter partial or incomplete data – a circumstance leading to significant performance degradation. Factors that can contribute to data

incompleteness are, for instance, privacy concerns that require the omission of certain modalities; device and security constraints that may preclude comprehensive data capture; and unintentional data leakages that may result in incomplete datasets. As a practical example, missing modality caused by various clinical and social reasons is a common issue in real-world healthcare scenarios [36].

The missing modality problem impacts early and late fusion models differently. In early fusion models the main encoder, trained to jointly process information from multiple modalities, loses part of its input, disrupting the fine-grained multimodal interactions it used to rely on and reducing performance. In late fusion models, the encoder for the missing modality becomes unusable, though the other encoders remain unaffected. The final layer, which combines feature vectors, receives an out-of-distribution input, leading to performance degradation.

Such problems are deeply different and need different solutions. Since mitigating the missing modality problem for an entire deep encoder is much more challenging than the final shallow aggregator, we focus on the former case in our analysis. In particular, following Ma *et al.* [23] and Lee *et al.* [16], we carry on our analysis focusing on Transformer-based early fusion models, adopting the ViLT architecture for our experiments [14]. Lee *et al.* [16] have been among the first to investigate the missing modality problem and proposed a realistic experimental setting in which the missing modality cannot be known apriori and the data leakages can occur both during training and test phases. The same approach copes with the missing modality problem via prompt learning (Fig. 1), an effective and efficient transfer-learning technique that avoids the fine-tuning of the whole architecture [3, 7, 11, 17]. Their approach requires training a different pool of prompts for each missing modality scenario. As the number of missing modality scenarios scales exponentially with the number of possibly missing modalities[1], the adoption of such methodology is unfeasible for systems having several modalities. Moreover, robustness concerns can also be raised in the simplest case with two modalities. Indeed, if during the training phase the three prompt pools are not trained evenly, some of them may encounter overfitting or underfitting, negatively affecting performance during inference. Most importantly, such prompts have limited expressive and adaptation power, as they need to be shared across all possible input samples.

These shortcomings motivate us to devise a prompting methodology independent of the number of missing modality scenarios. In particular, we propose a novel prompt-learning module that can generate sample-specific prompts explicitly conditioned on the semantics of input samples, rather than solely relying on the missing modality case. Through the proposed prompt module, termed SCP, we are able to exploit a vast spectrum of prompts alongside a limited number of learnable parameters, ending up with a flexible and robust solution that automatically adapts multimodal Transformers to datasets with different semantics and number of modalities. Also, to ensure proper transfer learning, we integrate our semantically-conditioned prompts with agnostic prompts. Experimentally, we validate our solution across a range of multimodal datasets, namely MM-IMDb [2], Food-101 [32], and Hateful Memes [13], and demonstrate its effectiveness in comparison with previous methods and baselines.

Contributions. To sum up, our contributions are as follows:

- We propose a semantically-conditioned prompting module that leverages the semantic representation of the available modalities to generate prompts tailored for each specific sample.
- We show that integrating agnostic prompts with semantically conditioned prompts promotes Transformer adaptation to both task and missing mode scenarios;
- Experimentally, we demonstrate that SCP is more resilient to extreme cases of complete/incomplete balance ratio, making it more reliable in real scenarios;
- Through t-SNE [31] visualizations, we show how the agnostic prompts used by our SCP reorganize themselves in as many clusters equal to the missing modality scenarios, preventing the burden of instantiating a pool of prompts for each of them, and that our semantically-conditioned prompts effectively organize themselves in different semantic levels.

2. Related Work

Multimodal Learning and Missing Modalities. The integration of heterogeneous data streams in Deep Learning models is a demanding problem in multimodal learning [6, 9, 14, 18]. Among them, the *missing modality problem*, wherein one or more modalities may be absent during inference or even training, is one of the most challenging.

Contemporary studies [16, 23, 24, 38] have been directed towards the development of multimodal frameworks capable of handling datasets with absent modalities. The SMIL approach [24] is introduced to infer the latent features of data with incomplete modalities using Bayesian meta-learning. Zeng *et al.* [37] has designed a tag-encoding mechanism that aids in the training of Transformer encoders to cope with absent modalities. The MMIN method [38] deduces the missing modality representation leveraging a unified multimodal representation from the remaining available modalities through cross-modality imagination with stacked residual autoencoders [29]. Furthermore, Ma *et al.* [23] delve into the resilience of multimodal Transformers when faced with missing modalities, improving their robustness by automatically searching for an optimal fusion

[1] Given M modalities, the number of missing modality scenarios is $2^M - 1$.

strategy. Lately, Lee *et al.* [16] exploits pools of learnable prompts for each possible *missing modality scenario* to mitigate missing modality issues. Despite its effectiveness, we recognize two main flaws. Firstly, it requires a number of prompt pools equal to the number of different scenarios that, if not equally trained, can undermine model robustness. Secondly, it only conditions the prompting with respect to the missing modality scenario, providing the same prompt regardless of the semantics of the input.

Prompt Learning. It marks a strategic advancement in transfer learning, providing a resource-efficient alternative to the computational demands of fine-tuning large-scale Transformer models. It adapts task-specific input prompts to leverage the pre-existing knowledge of the model without extensive retraining. Prompts can be either expert-designed [7] or learned autonomously by the model, with techniques like prompt-tuning [17] and prefix-tuning [19] being prominent examples. These strategies enhance the generalizability of Transformer models to downstream tasks, even in few-shot or zero-shot settings.

In vision, visual prompts [3, 11] modify Transformers for specific tasks, with techniques like L2P [35] and DualPrompt [34] addressing continual learning. CoOp [39] adapts CLIP-like models for image recognition, while Frozen [30] trains visual encoders to generate prefix prompts for guiding pre-trained language models. MaPLe [12] applies prompts across visual and text encoders to improve cross-domain representation. These studies inspired us to integrate prompt learning into our solution.

3. Proposed Method

Our method addresses the challenge of missing modalities in the field of multimodal learning. Consistently with the current state-of-the-art benchmarks and previous literature [16], our model is evaluated on two different modalities: text and image. In this setting, three missing modality scenarios are possible: "missing image", "missing text", and "complete" (*i.e.*, when both modalities are available). Our experimental setting also considers the case where the missing modality issue may occur both in the training and testing phases, mirroring a realistic scenario.

Within this scope, our main contribution consists of a novel prompting methodology, called SCP, that supports a multimodal Transformer in mitigating the missing modality problem. As opposed to the state-of-the-art, SCP is independent of the number of missing modality scenarios and aims at generating input prompts by conditioning them with respect to the semantics of the input samples rather than solely relying on the missing modality case. Thanks to this semantic conditioning, SCP exploits a vast spectrum of possible prompts, which are more consistent with the available input modalities, alongside a limited number of learnable parameters, enhancing flexibility and robustness.

3.1. Preliminaries

Problem Statement. We consider a multimodal dataset comprising $M = 2$ modalities, m_1 and m_2, exemplified by images and text, recalling that the number of missing modality scenarios for a generic dataset D is $2^M - 1$. Hence, for a bimodal dataset we have three scenarios, defined as: $D^c = \{(x_i^{m_1}, x_i^{m_2}, y_i)\}_i$ for "complete" data pairs, $D^{m1} = \{(x_i^{m1}, y_i)\}_i$ and $D^{m2} = \{(x_i^{m_2}, y_i)\}_i$ for modality-incomplete data, where y_i represents ground-truth labels. The training data is a mixture of these subsets and is defined as $D = D^c \cup D^{m1} \cup D^{m2}$. To maintain input consistency, dummy inputs $\tilde{x}^{m1}$, $\tilde{x}^{m2}$ (such as empty strings or black images) are used to represent the absence of modalities, creating reformed subsets $\tilde{D}^{m1}$ and $\tilde{D}^{m2}$, defined as $\tilde{D}^{m1} = \{(x_i^{m1}, \tilde{x}^{m2}, y_i)\}_i$ and $\tilde{D}^{m2} = \{(\tilde{x}^{m1}, x_i^{m2}, y_i)\}_i$, respectively. Therefore, the final dataset can be formulated as $\tilde{D} = D^c \cup \tilde{D}^{m1} \cup \tilde{D}^{m2}$. In our case, we will refer to $\tilde{D}^{m1}$ and $\tilde{D}^{m2}$ as image-only (or "missing text") and text-only (or "missing image") scenarios.

Furthermore, it is worth mentioning that in the remainder of the paper, we will use the term *missing modality scenario* to describe the state of a specific data sample. To define the set of all the possible missing modality scenarios eligible in our experiments, we use the term *missing modality cases*[2]. Specifically, three missing modality cases can arise in our context: "missing text", "missing image", and "missing both", respectively defined as $\tilde{D^{mt}} = D^c \cup \tilde{D}^{m1}$, $\tilde{D^{mi}} = D^c \cup \tilde{D}^{m2}$, and $\tilde{D^{mb}} = D^c \cup \tilde{D}^{m1} \cup \tilde{D}^{m2}$.

Backbone. Considering its widespread application in various multimodal learning tasks, we employ a pre-trained multimodal Transformer ViLT [14] as reference backbone. ViLT is a Vision Transformer (ViT) [8], which can take as input the concatenation of textual and visual tokens and which has shown proficient capabilities alongside faster inference times with respect to its competitors.

As ViLT inherits its structure from ViT, it makes use of a `[CLS]` token and a pooler layer. `[CLS]` is a learnable token prepended to each input sequence, and is designed to summarize the information of the whole sequence at each layer of the Transformer. This learnable token integrates the information from all the other tokens thanks to the Multi-head Self-Attention (MSA) operator that allows information exchange between all the possible couples of tokens and creates a hidden state representation of the entire sequence. We will refer to the hidden state of the `[CLS]` token at the i-th layer as $\texttt{[CLS]}^i$.

The pooler layer, instead, extracts the hidden state of the `[CLS]` token from the last layer and takes a linear projection of it, followed by a $\tanh$ activation. For the sake of solving our multimodal tasks, the output of the pooler is fed

[2]We refer to the supplementary material for a graphic representation of the considered scenarios.

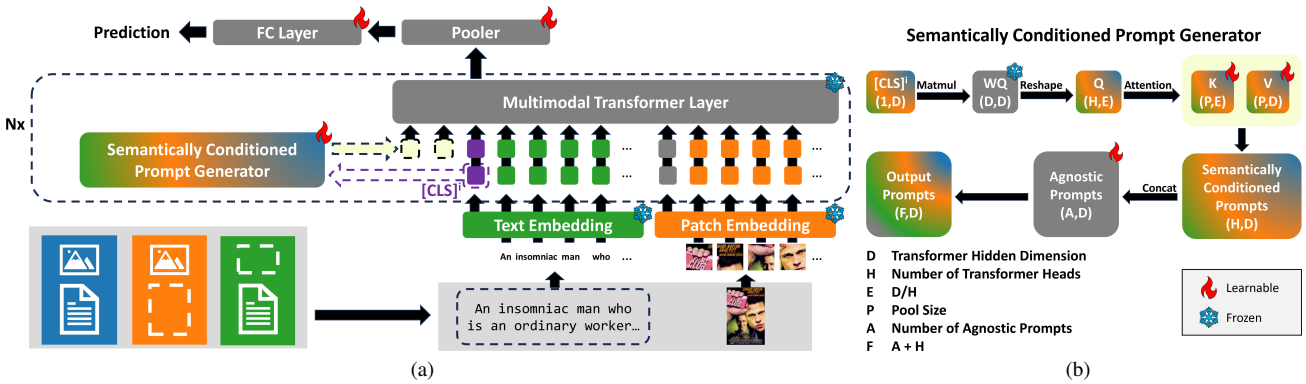


Figure 2. (a) Prompt-based multimodal framework. Different types of missing modality scenarios characterizing the input data stream are depicted at the bottom-left. On the bottom-right corner, text and image are preprocessed before being fed into the Transformer-based architecture. Next, the hidden states of the [CLS] for each selected layer (not the [CLS] itself), defined as $\texttt{[CLS]}^i$, are fed into the *semantically conditioned prompt generator*. The generated prompts are prepended to the rest of the sequence, and the entire sequence is fed into a ViLT encoder layer. The last $\texttt{[CLS]}^i$ of the ViLT output sequence is extracted and fed into the Pooler. The pooled output is finally fed into fully-connected layers to predict the final output. (b) Internal architecture of our semantically conditioned prompt generator.

into a task-specific layer composed of a fully connected network, mapping the pooled representation into a space that matches the number of task classes. Considering that our methodology employs prompt learning, we will keep the parameters of ViLT frozen and update only the weights of the prompts, the pooling layer, and task-specific layers.

Prompt Integration. Given a pre-trained multimodal Transformer f_θ with N consecutive MSA layers, we represent the input features for the i-th MSA layer as $\boldsymbol{h}^i \in \mathbb{R}^{L\times d}$; $i = 1, 2, \ldots, N$, where L is the input length and d is the embedding dimension. Specifically, $\boldsymbol{h}^1$ is the concatenation of the text and visual tokens, obtained using the text tokenizer and visual embedder from the raw inputs. Moreover, additional prompts $\boldsymbol{p}^i \in \mathbb{R}^{L_p\times d}$ can be concatenated to the input before every i-th encoder layer of the ViLT architecture, where L_p is the length of added prompts. These prompts are concatenated along the sequence-length dimension with the input to generate augmented features $\boldsymbol{h}_p^i$, as

$$\boldsymbol{h}_{\boldsymbol{p}}^i = [\boldsymbol{p}^i; \boldsymbol{h}^i]. \tag{1}$$

Training Objective. As anticipated, we keep all the Transformer parameters θ frozen except for the task-specific layers θ_t. We identify as θ_p the parameters of the additional prompts. Hence, the overall objective can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{task}}(x_i^{m1}, x_i^{m2}, \theta_t, \theta_p), \tag{2}$$

where $(x_i^{m1}, x_i^{m2}) \in \tilde{D}$ represents the multimodal input pair subject to missing modality issues, and $\mathcal{L}_{\text{task}}$ symbolizes the task-specific multimodal objective.

3.2. Semantically Conditioned Prompts

Our proposed module, termed SCP, consists in a novel prompting methodology that assists a multimodal Transformer in mitigating the missing modality problem (Fig. 2a). SCP generates tailored prompts for each data sample employing a novel semantically conditioned prompt generator (see Fig. 2b), with the objective of providing prompts that are more consistent with the available input modalities. To achieve this, SCP needs to capture the semantics of the available input modalities.

Considering that we are operating inside a multimodal Transformer that makes use of the [CLS] token, we exploit its hidden states $\texttt{[CLS]}^i \in \mathbb{R}^{1\times d}$ at each layer as a representation of the semantics of the input sample. In particular, SCP leverages the information contained in $\texttt{[CLS]}^i$ to query a "memory bank" composed of learnable key-value pairs, exploiting a conventional attention mechanism, thus effectively generating semantically-conditioned prompts. Remarkably, the various $\texttt{[CLS]}^i$ are the only information that is always present in this experimental setting, regardless of the number of input tokens or modalities, and the only requirement of our methodology. Overall, this makes SCP independent of the number of available tokens or modalities and enhances its flexibility and robustness.

The first step of the prompt generation process for a generic input h^i requires the creation of query vectors from the hidden states $\texttt{[CLS]}^i$, exploiting the pre-trained query projection matrix $\boldsymbol{W}_Q^i \in \mathbb{R}^{d\times d}$ from the i-th MSA layer of the architecture. The W_Q^i matrix is a component of the i-th MSA layer of ViLT used for creating a number of query vectors equal to the number of Transformer heads n_h for each input token. Specifically, given a generic token $\boldsymbol{t} \in \mathbb{R}^{1\times d}$, $\boldsymbol{q} = \boldsymbol{t}\boldsymbol{W}_Q \in \mathbb{R}^{1\times d}$ is a vector that contains n_h queries of size d/n_h. To extract such queries, $\boldsymbol{q}$ must be splitted in n_h subvectors that can be concatenated obtaining $\boldsymbol{Q} \in \mathbb{R}^{n_h\times \frac{d}{n_h}}$. We refer to this combination of splitting and concatenation as the *reshape* operator. Accordingly, we make the same operations by projecting each $\texttt{[CLS]}^i$

through $\boldsymbol{W}_Q^i$, obtaining our queries:

$$\boldsymbol{Q}^i = \text{reshape}\left(\texttt{[CLS]}^i \boldsymbol{W}_Q^i\right) \in \mathbb{R}^{n_h \times \frac{d}{n_h}}. \quad (3)$$

In the second step, these queries are employed in a classic attention mechanism, where they are allowed to pay attention to a pool of learnable keys $\boldsymbol{K} \in \mathbb{R}^{P \times \frac{d}{n_h}}$, with P being the pool size. The attention scores, defined as $\text{scores} = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^\intercal}{\sqrt{d}})$ are used to retrieve learnable values $\boldsymbol{V}$, defined as $\boldsymbol{V} \in \mathbb{R}^{P \times d}$ through matrix multiplication. Finally, the semantically conditioned prompts are computed as

$$\boldsymbol{s}^i = \text{softmax}\left(\frac{\boldsymbol{Q}^i(\boldsymbol{K}^i)^\intercal}{\sqrt{d}}\right) V^i \in \mathbb{R}^{n_h \times d}. \quad (4)$$

Hence, each prompt is formed as a linear combination of values, contingent upon their alignment with the keys. Consequently, each data sample receives customized prompts that depend on the semantics captured by $\texttt{[CLS]}^i$.

Furthermore, to ensure model stability and offer a pathway for potential fine-tuning, we concatenate a few agnostic prompts $\boldsymbol{a}^i \in \mathbb{R}^{L_a \times d}$, where L_a is the length of agnostic prompts, alongside with those generated by our module. These additional prompts are just learnable vectors but can be crucial to improve the model stability since they are data-independent, while the generator module alone might lead to fluctuations during training. Finally, the prompts $\boldsymbol{p}^i = [\boldsymbol{a}^i; \boldsymbol{s}^i]$ are concatenated to the main input sequence.

Agnostic Prompts. The agnostic prompts have been designed to question whether having a specialized prompt for each case of missing modality scenario is worth it [16], or if a generic pool of prompts can adapt itself to different scenarios. From a practical point of view, the agnostic prompting is trivial, consisting of prepending a pool of learnable prompts to the input sequence regardless of the missing modality scenario for a predetermined subset of ViLT layers as shown in Eq. (1).

4. Experimental Results

We follow the experimental protocol defined by Lee *et al.* [16] and report experiments across a range of multimodal datasets, namely MM-IMDb [2], UPMC Food-101 [32], and Hateful Memes [13].

MM-IMDb [2] serves for movie genre classification, incorporating both image and text modalities. Given the multi-genre nature of movies, this dataset poses a multi-label classification challenge. The task involves predicting the set of genres a movie belongs to, utilizing either the image (movie poster), the text (movie plot), or a combination of both.

UPMC Food-101 [32] is constructed for the task of food classification, including both image and text modalities. It includes noisy image-text pairs retrieved from Google Image Search and aligns with the category structure of the ETHZ Food-101 dataset [5].

Hateful Memes [13] requires the identification of hate speech within memes using both image and text modalities. It is deliberately structured to foil unimodal models by introducing "benign confounders" thus necessitating effective multimodal analysis to achieve accurate classification.

Metrics. To assess the performance on these diverse tasks, we employ task-appropriate metrics for each dataset. For the multi-label classification on MM-IMDb, the F1-Macro score is used, providing a balanced measure of the model performance across multiple genres. In the case of UPMC Food-101, we utilize classification accuracy as the metric. For Hateful Memes, the evaluation score is the Area Under the Receiver Operating Characteristic Curve (AUROC).

4.1. Implementation Details

Input. Following [14], we resize images such that the shorter side is 384 pixels, and the longer side does not exceed 640 pixels, maintaining the original aspect ratio. Consistent with [8], images are divided into patches of size 32×32. If the image modality is missing, we use a dummy image composed of pixels with values set to one, indicated as $\tilde{x}_{m1}$. For the text modality, we employ the `bert-base-uncased` tokenizer to process the textual input. In the absence of text, an empty string is used as a dummy input, denoted as $\tilde{x}_{m2}$. The maximum length for text inputs is set to $1,024$ for MM-IMDb, 512 for UPMC Food-101, and 128 for Hateful Memes.

Multimodal Backbone. In our experiments, we adopt the ViLT version pre-trained with image text matching and masked language modeling objectives on MSCOCO [21], Visual Genome [15], CC3M [28], and SBU [25].

Model Training. As mentioned, in our model configuration the ViLT backbone parameters are kept frozen, while the training is confined to the learnable prompts and parameters related to the downstream tasks, specifically the pooler and the task-specific classifiers. In particular, our ViLT counts 12 layers and 12 heads per layer. Hence, the number of semantically conditioned prompts generated is 12, and we set the number of SCP agnostic prompts to 4 so that the length L_p of learnable prompts is equal to 16. Moreover, we set the number of learnable keys and values of SCP equal to 32. We use agnostic prompts only from the 1st to 6th layer and our SCP module from the 7th to the 12th layer, configured as discussed above. We use the AdamW optimizer [22] with a base learning rate of 1×10^{-2} and a weight decay of 2×10^{-2}. The learning rate undergoes a warm-up period for the initial 10% of the training steps and subsequently decays linearly to zero.

Setting of Missing Modality. Our work addresses a generalized missing modality scenario applicable to both the training and testing phases. Hence, each data sample within a modality can be subject to leakages, defining the miss-

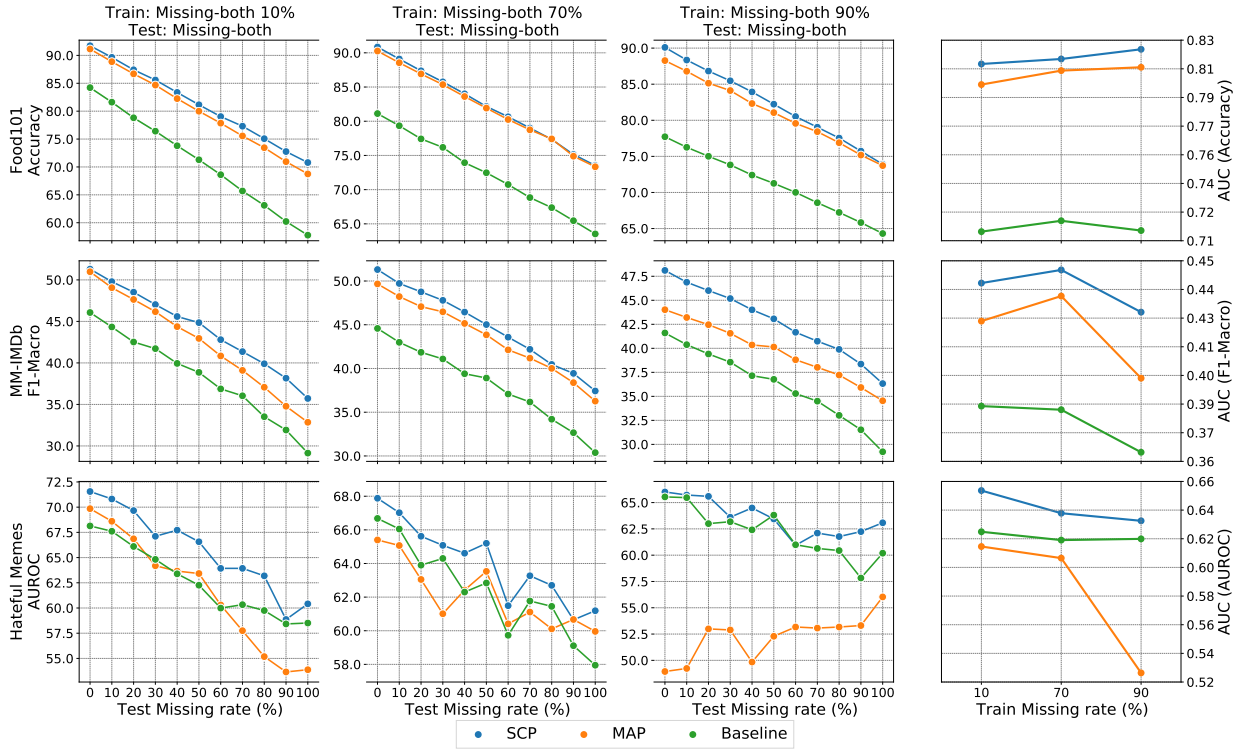


Figure 3. Robustness to different train/test missing rates of SCP, MAP, and Baseline on Food101, MM-IMDb, and Hateful Memes.

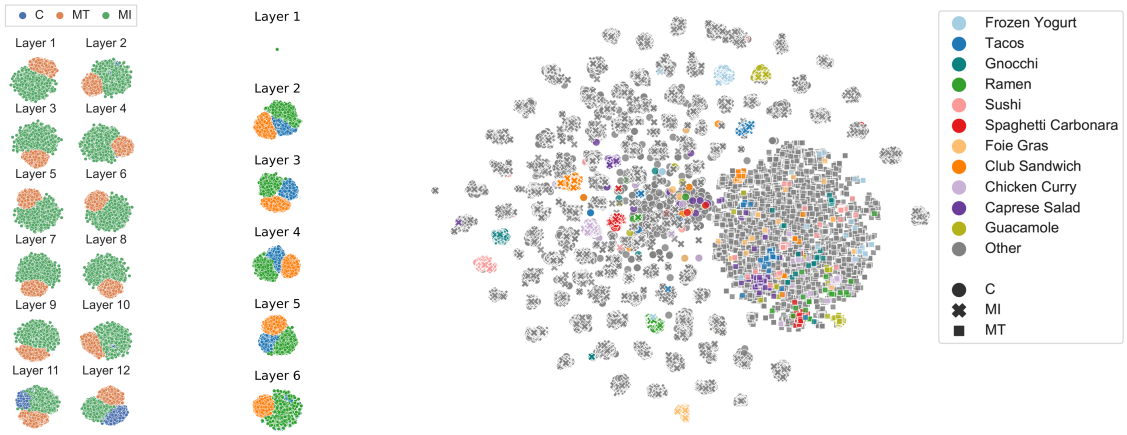
ing rate $\eta\%$ as the proportion of incomplete data within the entire dataset. In the missing-text (or missing-image) case with a missing rate $\eta\%$, the dataset consists of $\eta\%$ image-only (or text-only) data and $(1-\eta)\%$ of complete data. For the missing-both case, the data is partitioned into $\frac{\eta}{2}\%$ text-only, $\frac{\eta}{2}\%$ image-only, and $(1-\eta)\%$ complete data. This partitioning scheme extends to tasks with M modalities, resulting in $(\frac{\eta}{2^M-2})\%$ incomplete data for each missing modality scenario and $(1-\eta)\%$ complete data.

4.2. Main Results

In our experiments, we compare our SCP module with the Missing-Aware Prompts (MAP) proposed in [16], which employs the same backbone and experimental settings of our approach. Additionally, we include the results of a baseline model, which does not employ any prompts and relies solely on fine-tuning the pooler and fully-connected layers to establish a reference for performance gains.

Robustness to Different Missing Rates. Well-designed multimodal architectures should be robust to any combination of missing modality rates, both in training and testing phases, as in real-world scenarios such rates may vary over time due to several causes. For this reason, we design an experiment to evaluate the robustness of our proposal with respect to our main competitor MAP [16] and the baseline model. Specifically, we train the models with three different train missing rates (*i.e.*, 10%, 70%, and 90%) under the missing-both case. Then, we test the models at different missing rates, varying them in a range from 0% to 100% with a step of 10% again under the missing-both case. We chose these values because they represent three extreme cases in our experimental setting. In particular, with a train missing rate of 10%, the models have few occasions to learn how to mitigate the missing modality problem, thus we expect the models to gain performances in the cases where few data are missing while suffering significant losses with severe missing rates. The opposite case is the train with a missing rate of 90%, where the models should have learned well how to mitigate the missing modality problem but are showing lower performances than usual in modality-complete scenarios. Finally, the 70% missing rate case is the most ideal. Hence, in this case, the modality complete scenario occurs in 30% of time, while missing-text and missing-image occur in the remaining 70% of the time evenly, each for 35%. Indeed, we should expect the models to be more robust in general. In addition, these patterns may vary due to the difficulty of the task and inherent dataset characteristics.

Results are shown in Fig. 3 for SCP, MAP, and the baseline on the Food101, MM-IMDb, and Hateful Memes datasets. 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. As the graph shows, SCP gains on average 1.5 AUC points against MAP on both Food101 and MM-IMDb in every experimental setting. Moreover, it is worth mention-



(a) (b) (c)

Figure 4. t-SNE visualizations of: (a) $\texttt{[CLS]}^i$ hidden states at the output of each layer of the pre-trained ViLT architecture. (b) attention paid by each $\texttt{[CLS]}^i$ to SCP's agnostic prompts. (c) attention weights of the semantically conditioned prompt generator module of SCP for 10 random classes of the Food101 test set, the remaining labels have been aggregated into 'Other' to enhance visual clarity.

ing that with the Hateful Memes dataset, MAP is incapable of surpassing the baseline model, while our SCP is more than 3 AUC points above. As expected, with the balanced case (*i.e.*, with a train missing rate of 70%) the gap with the competitor becomes narrower. Overall, these results clearly show the benefits of using semantically conditioned prompts to effectively handle different missing modality scenarios and rates, demonstrating consistent improvements compared to the competitor and baseline.

Attention Visualizations with t-SNE. To understand the effectiveness of the agnostic and semantically conditioned prompts, we leverage the t-SNE method [31]. In this case, all experiments are conducted on the test set of Food101 to demonstrate the generalizability of our considerations on unseen data. A preliminary analysis aims to understand whether the pre-trained ViLT is able to identify missing modality scenarios without requiring external information and exploiting only the hidden states $\texttt{[CLS]}^i$. For this purpose, Fig. 4a provides a representation of the $\texttt{[CLS]}^i$ drawn from the output of each layer constituting the pre-trained ViLT architecture for each test sample. In the earlier stages, the modality-complete (blue) almost overlaps with missing-image (green), while missing-text (orange) is clearly distinguishable from the others. This is an expected behavior since text carries most of the semantics at earlier stages of ViLT. Instead, as we consider deeper encoder layers, the visual features gain importance and, when facing a modality-complete scenario, ViLT can exploit them to capture different semantics with respect to text-only inputs.

Fig. 4b showcases that a pool of agnostic prompts can automatically adjust itself to tackle different modality cases without any manual adjustment. In this case, t-SNE is used to represent the attention paid by the [CLS] token and subsequent hidden states $\texttt{[CLS]}^i$ to the agnostic prompts of the ViLT architecture trained using our SCP prompting strategy for each test sample. 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. Also, 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.

Finally, to demonstrate that SCP is capable of generating tailored prompts conditioned on the semantics of each sample, Fig. 4c is also provided. For the sake of ensuring an easier visualization, we decided to randomly sample 10 classes from the Food101 dataset, while retaining the visualization efficacy without loss of generalization. In this case, t-SNE is used to represent the attention weights of the semantically conditioned prompt generator of SCP of each test sample. 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 sub-clusters, 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. As shown in Fig. 4a, ViLT overrelies on text semantics for multimodal tasks, indicating it carries the most predictive information. Thus, in the missing-text scenario, the $\texttt{[CLS]}^i$ token exhibits weaker semantics, adversely affecting the attention mechanism of SCP.

4.3. Ablation Studies

The ablation studies conducted in our research serve as a rigorous examination of the discussed prompting strategies.

Agnostic Prompt Effectiveness. To determine if agnostic prompts can adapt to different missing modality scenarios

Table 1. Average performance over three runs by changing the random seed. All the experiments have missing rate η equal to 70%. The percentage of availability for each modality can be found in the second column. Best results in **bold**.

Dataset	Train/Test *Text*	Train/Test *Image*	Baseline	MAP [16]	AP (Ours)	SCP (Ours)
MM-IMDb (F1-Macro)	30%	100%	33.97	37.05	37.04	**37.19**
	100%	30%	37.81	46.26	47.58	**48.16**
	65%	65%	36.22	41.37	41.46	**41.67**
Food101 (Accuracy)	30%	100%	66.19	73.21	73.52	**73.97**
	100%	30%	76.66	86.44	86.33	**86.56**
	65%	65%	69.14	78.51	78.34	**78.99**
Hateful Memes (AUROC)	30%	100%	59.26	59.50	58.86	**60.07**
	100%	30%	63.02	62.74	64.13	**64.26**
	65%	65%	62.35	61.82	62.36	**62.90**

quantitatively, we train a model that solely harnesses agnostic prompts to mitigate the missing modality problem. We compare it with MAP [16], the baseline model, and our SCP. Results are presented in Tab. 1, covering each dataset, missing rate, and modality case. Each value in this table represents the average metric obtained by repeating each experiment three times, varying only the random seed generator to get a better performance evaluation, as opposed to MAP, which reports single-run experiments. Tab. 1 shows that our proposed method outperforms MAP on all considered datasets and missing modality cases and, intriguingly, agnostic prompts mirror the performance of our competitor. This further confirms the need to use semantically conditioned prompts to improve the final performance.

Missing Modality Aware Prompt Effectiveness. To understand whether missing modality aware prompt pools have some similarities between each other or if they are even interchangeable, we train the missing modality aware prompts accordingly to [16], reproducing the experiment on the Food101 [32] dataset with the missing rate η equal to 70% (35% for both text and image modalities). Then, during inference, we force the model to always take the prompts learned for a specific missing modality scenario, and we do it for all the cases: modality-complete, missing-text, and missing-image. Results are reported in Tab. 2. As we can see, when the model is forced to always use the prompts learned in the modality-complete scenario, the performances drop from 78.51% to 75.95%. Anyway, the performances are far above the baseline, which scores 69.14%. Hence, we can conclude that these prompts can learn patterns that can be useful regardless of the missing modality case, highlighting the fact that agnostic prompting can be a valuable solution. For the other two cases, performance drops significantly under the baseline, but this can be acceptable because prompts tailored for specific missing modality cases are unlikely to be useful for other cases.

Balance Between Agnostic and SCP Prompts. We conduct experiments to determine the optimal balance between agnostic and semantically conditioned prompts. As discussed, SCP performs better with agnostic prompts. Since `[CLS]` tokens are identical at the first layer, conditioning on them is ineffective, and relying solely on generated prompts may cause instability during fine-tuning. To address this, SCP uses only a pool of agnostic prompts in the first layer and combines agnostic and generated prompts in subsequent layers. We test SCP with and without concatenating agnostic prompts and evaluate the balance between the two types. Results in Tab. 3 show that a correct balance of agnostic prompts and SCP prompts improve performance. While agnostic prompts are more effective in early layers, SCP becomes essential in deeper layers.

Table 2. Ablation study on modality-complete (C), missing-image (MI), and missing-text (MT) prompts. "Available Prompts" indicates whether the model can choose between none, all the prompts (*i.e.*, C, MI, MT), or always the same single prompt pool to be prepended to the input sequence.

Dataset	Training *Text*	Training *Image*	Available Prompts	Accuracy
Food101	65%	65%	-	69.25
			MT	42.21
			MI	61.31
			C	75.95
			C, MI, MT	78.51

Table 3. Ablation study on the balance between agnostic and semantically conditioned prompts. Best results in **bold**.

Dataset	Training *Text*	Training *Image*	# Pools *Agnostic*	# Pools SCP	#AP	Accuracy
Food101	65%	65%	6	6	0	78.28
			6	6	4	**78.99**
Food101	65%	65%	1	11	4	78.79
			6	6	4	**78.99**
			9	3	4	78.92

5. Conclusion

In this work, we propose a novel prompting methodology, called SCP, that aims to mitigate the missing modality problem in large-scale multimodal Transformers such as ViLT. The proposed method generates semantic conditioned prompts leveraging the available input modalities information, creating ad-hoc prompts for each sample and missing modality scenario. SCP achieves the best performances compared to the previous state of the art on three datasets for visual recognition. Moreover, through t-SNE visualization, we show the capability of the proposed approach to generate prompts tailored for each semantic level.

Acknowledgements. This work has been supported by the PNRR project "Fit for Medical Robotics (Fit4MedRob)" (CUP B53C22006950001), funded by the Italian Ministry of University and Research, the PRIN 2022-PNRR project "MUCES" (CUP E53D23016290001), funded by the European Union - NextGenerationEU, and by the University of Modena and Reggio Emilia and Fondazione di Modena, through the FAR 2023 and FARD-2024 funds ("Fondo di Ateneo per la Ricerca").

References

[1] Hassan Akbari, Yin Yuan, Wei Feng, Wei Hu, Yanjun Wang, M Javad Roshtkhari, and Greg Mori. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *Advances in Neural Information Processing Systems*, 2021. 1

[2] 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, 5

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring Visual Prompts for Adapting Large-Scale Models. *arXiv preprint arXiv:2203.17274*, 2022. 2, 3

[4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 1

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–Mining Discriminative Components with Random Forests. In *Proceedings of the European Conference on Computer Vision*, 2014. 5

[6] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-End Referring Video Object Segmentation with Multimodal Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 2020. 2, 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 3, 5

[9] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-Modal Fusion Transformer for Video Retrieval. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *Proceedings of the European Conference on Computer Vision*, 2022. 2, 3

[12] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MAPLE: Multi-modal Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[13] 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, 5

[14] Wonjae Kim, Bokyung Son, and Ildoo Kim. VILT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 1, 2, 3, 5

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 5

[16] 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. 1, 2, 3, 5, 6, 8

[17] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021. 2, 3

[18] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*, 2021. 2

[19] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021. 3

[20] Paul Pu Liang and Louis-Philippe Morency. Tutorial on Multimodal Machine Learning: Principles, Challenges, and Open Questions. In *Proceedings of the ACM International Conference on Multimodal Interaction*, 2023. 1

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, 2014. 5

[22] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*, 2018. 5

[23] 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*, 2022. 1, 2

[24] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal Learning with Severely Missing Modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2

[25] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, 2011. 5

[26] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 1

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 1

[28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018. 5

[29] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[30] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems*, 2021. 3

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 2, 7

[32] 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, 5, 8

[33] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. TransModality: An End2End Fusion Method with Transformer for Multimodal Sentiment Analysis. In *Proceedings of the Web Conference*, 2020. 1

[34] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning. In *Proceedings of the European Conference on Computer Vision*, 2022. 3

[35] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning To Prompt for Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[36] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal Patient Representation Learning with Missing Modalities and Labels. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[37] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. 2

[38] Jinming Zhao, Ruichen Li, and Qin Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2021. 2

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

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.

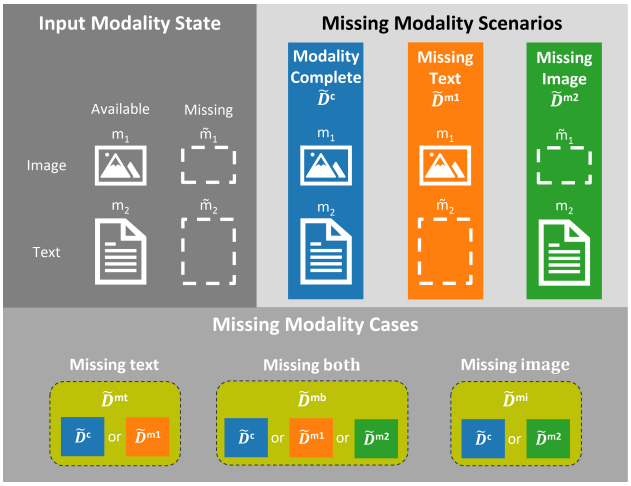


Figure 5. Visual diagram illustrating the potential input modality states (impacting each sample modality), missing modality scenarios (affecting each input sample), and missing modality case (impacting each experiment).

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	69.50	77.56	**78.54**
		image	77.65	87.22	**87.49**
MM-IMDb	AUC (F1-Macro)	text	36.80	39.71	**40.73**
		image	39.93	46.89	**49.31**
Hateful Memes	AUC (AUROC)	text	60.99	61.14	**61.43**
		image	64.56	60.18	**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%	71.08	80.01	**81.27**
		70%	71.74	80.87	**81.57**
		90%	71.15	81.07	**82.16**
MM-IMDb	AUC (F1-Macro)	10%	38.41	42.40	**44.18**
		70%	38.24	43.57	**44.80**
		90%	36.25	39.72	**42.82**
Hateful Memes	AUC (AUROC)	10%	62.93	61.54	**65.67**
		70%	62.01	60.69	**64.00**
		90%	62.10	52.21	**63.44**

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

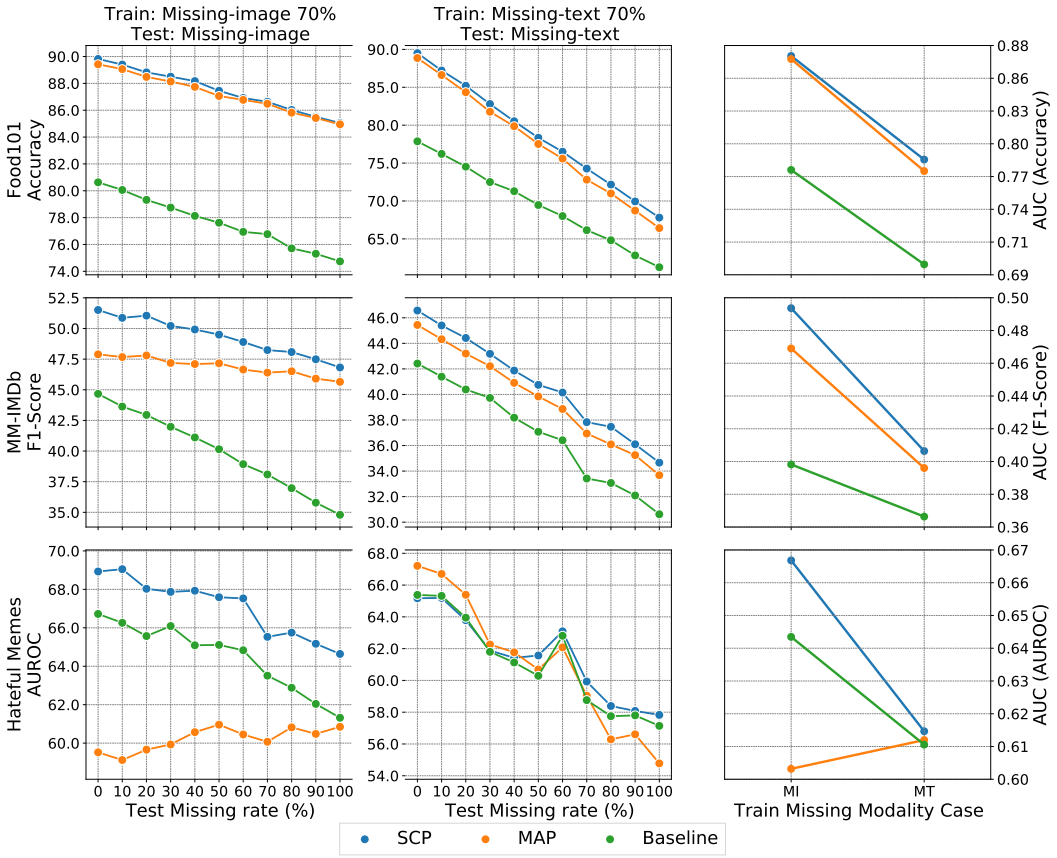


Figure 6. Robustness to different Train Missing Modality Cases and Test Missing Rates of SCP, MAP, and Baseline on Food101, MM-IMDb, and Hateful Memes. The Train Missing Rate is fixed at 70% for all the experiments.

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.

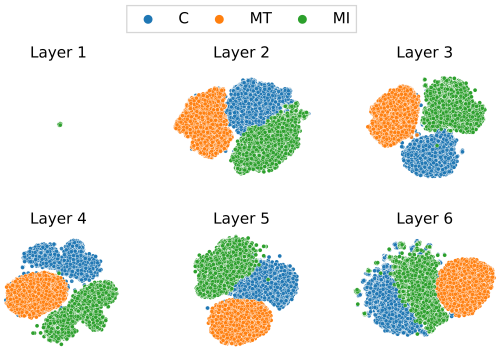


Figure 7. t-SNE visualization of the attention that the `[CLS]` and subsequent hidden states $\texttt{[CLS]}^i$ pay to the agnostic prompts for the first 6 layers of the ViLT architecture. Such ViLT architecture only harnesses agnostic prompts without SCP. In this way, the contribution of agnostic prompts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

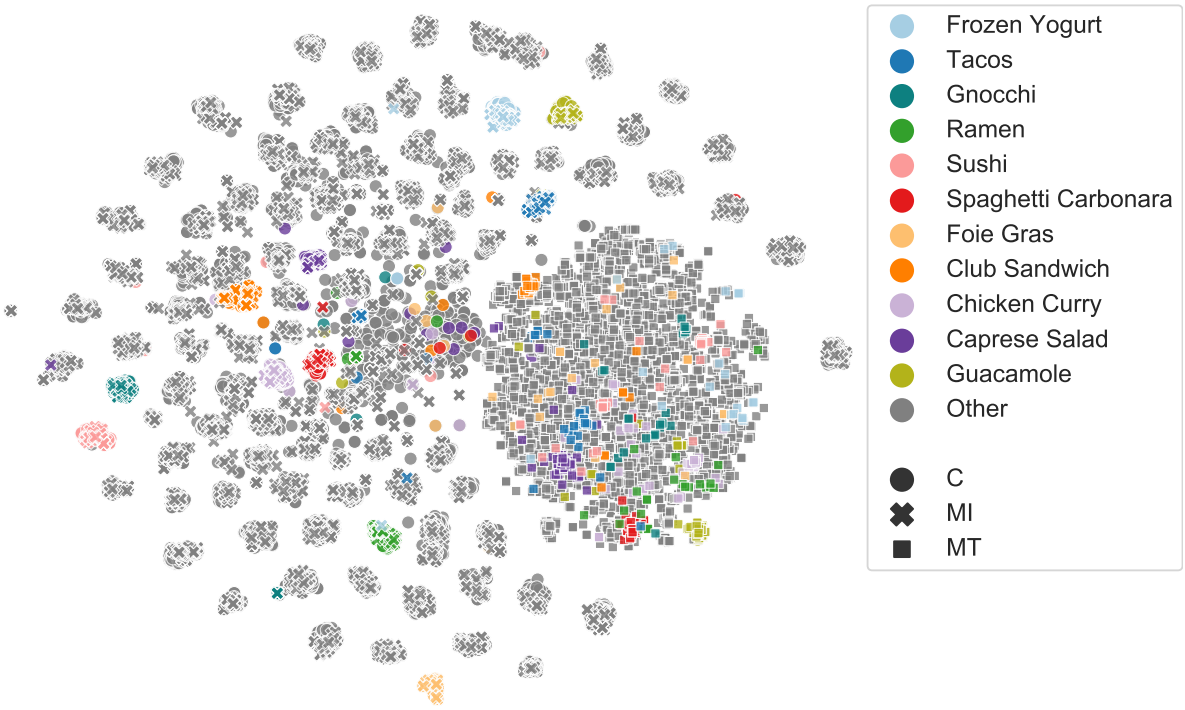


Figure 8. Attention weights of the semantically conditioned prompt generator module of SCP for the Food101 [4] test set. We provide only the annotation of the first 10 classes to reduce confusion in the plot.

[CLS] token and subsequent hidden states $[CLS]^i$ 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 $[CLS]^i$ representations to properly work. Indeed, the missing-text scenario leads to $[CLS]^i$ 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