Semantically Conditioned Prompts for Visual Recognition under Missing Modality Scenarios

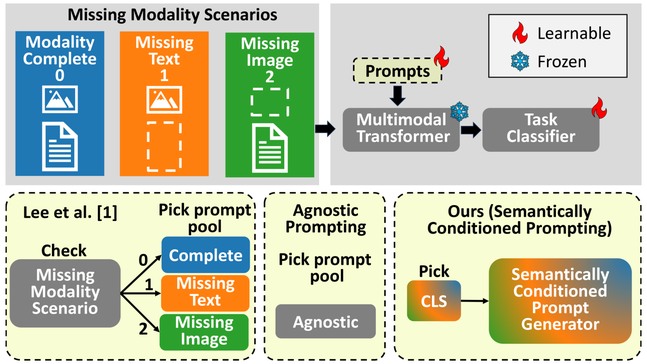
Vittorio Pipoli[1,2], Federico Bolelli[1], Sara Sarto[1], Marcella Cornia[1], Lorenzo Baraldi[1], Costantino Grana[1], Rita Cucchiara[1], Elisa Ficarra[1]

[1]University of Modena and Reggio Emilia, Italy {name.surname@unimore.it}; [2]University of Pisa, Italy {name.surname@phd.unipi.it}

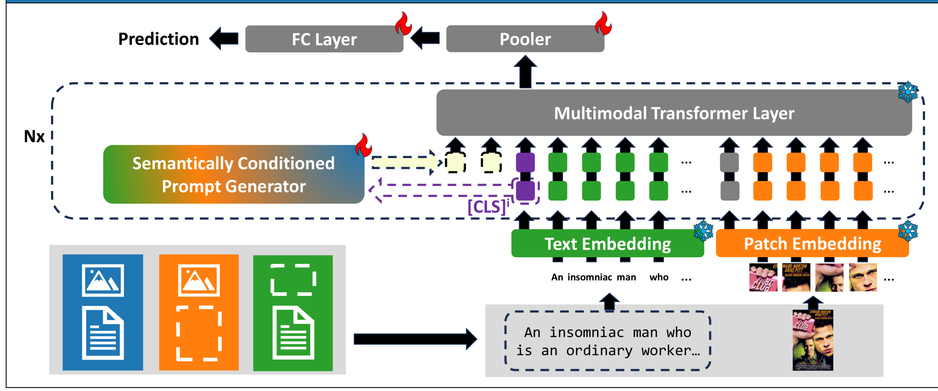
WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

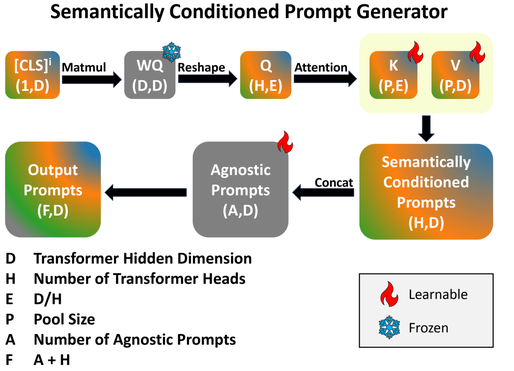
1 - Missing Modalities and Multimodal Transformers

Prompt learning has proven effective in mitigating **missing modalities** in **multimodal transformers**. This work enhances the SOTA by Lee et al. [1] with **semantically conditioned prompts** that dynamically adapt to input semantics and missing modality scenarios.



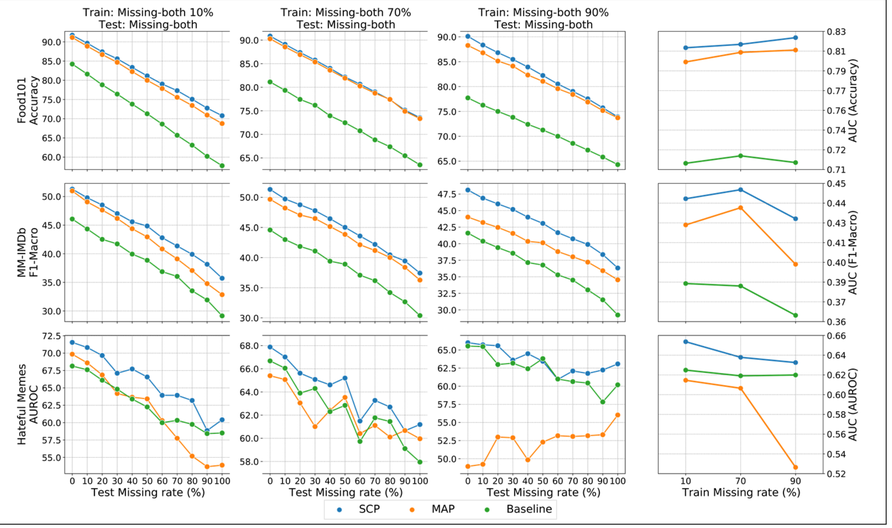
2 - Method: Semantically Conditioned Prompts for Multimodal Transformers



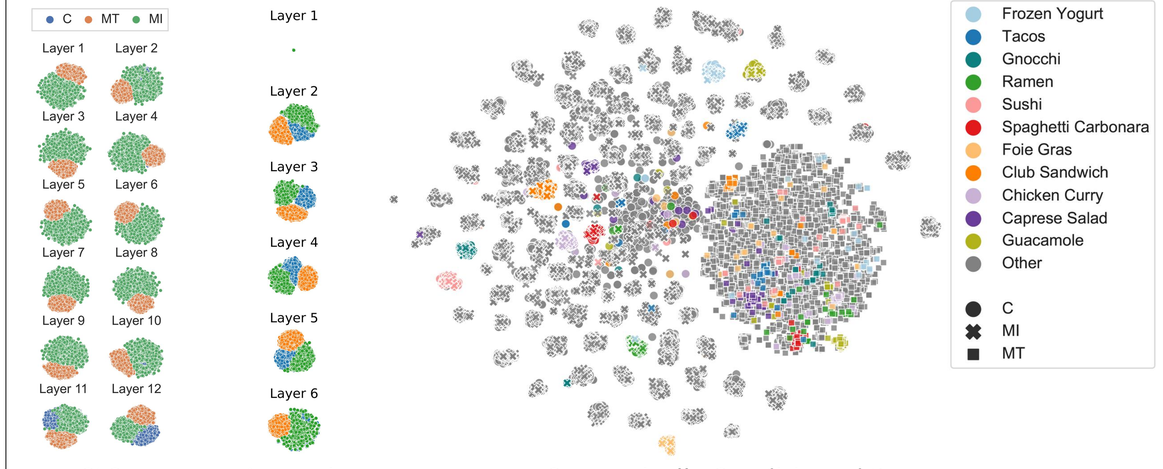


3 - Quantitative Results: Robustness to Different Missing Rates

We tested our proposed **SCP** against **MAP** (Lee et al.) and a **Baseline** (no prompts, only Pooler and FC Layer tuning) on three multimodal datasets: **Food-101** [2], **MM-IMDb** [3], and **Hateful Memes** [4]. We measured the **robustness to different missing rates** varying both train and test missing rates. **The rightmost column depicts the area under the curve** of the respective metric curves on the left. **SCP is the best performing model.**



4 - Visualization of CLS Tokens and Attention Weights (t-SNE)



[1] Lee et al., Multimodal Prompting with Missing Modalities for Visual Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.
[2] Wang et al., Recipe recognition with large mul-timodal food dataset. In Proceedings of the IEEE Inter-national Conference on Multimedia and Expo Workshops, 2015.
[3] Arevalo et al., Gated Multimodal Units for Informa-tion Fusion. In Proceedings of the International Conferenceon Learning Representations Workshops, 2017
[4] Kiela et al., The Hateful Memes Challenge: Detecting HateSpeech in Multimodal Memes. In Advances in Neural Infor-mation Processing Systems, 2020