Efficient Prompting for Continual Adaptation to Missing Modalities

Zirun Guo, Shulei Wang, Wang Lin, Weicai Yan, Yangyang Wu[*], Tao Jin
Zhejiang University
zrguo.cs@gmail.com

Abstract

Missing modality issues are common in real-world applications, arising from factors such as equipment failures and privacy concerns. When fine-tuning pre-trained models on downstream datasets with missing modalities, performance can degrade significantly. Current methods often aggregate various missing cases to train recovery modules or align multimodal features, resulting in suboptimal performance, high computational costs, and the risk of catastrophic forgetting in continual environments where data arrives sequentially. In this paper, we formulate the dynamic missing modality problem as a continual learning task and introduce the continual multimodal missing modality task. To address this challenge efficiently, we introduce three types of prompts: modality-specific, task-aware, and task-specific prompts. These prompts enable the model to learn intra-modality, inter-modality, intra-task, and inter-task features. Furthermore, we propose a contrastive task interaction strategy to explicitly learn prompts correlating different modalities. We conduct extensive experiments on three public datasets, where our method consistently outperforms state-of-the-art approaches.

1 Introduction

Pre-trained multimodal models have shown great potential in many applications (Radford et al., 2021; Li et al., 2023; Lin et al., 2024). When fine-tuning these pre-trained models on downstream tasks, missing modality issues often occur due to equipment failure, data corruption, privacy concerns, etc. Existing methods (Ma et al., 2021; Zhao et al., 2021; Lee et al., 2023; Guo et al., 2024c) address missing modality issues by reconstructing missing information or aligning multimodal features. However, both recovering missing features and aligning multimodal features are based on datasets containing various types of missing modality cases (Figure 1 (left)) to achieve robust performance. For example, recovering feature methods learn how to reconstruct a missing modality using the available modalities. Hence, it is expected that the dataset contains various types of missing cases to optimize the reconstruction modules.

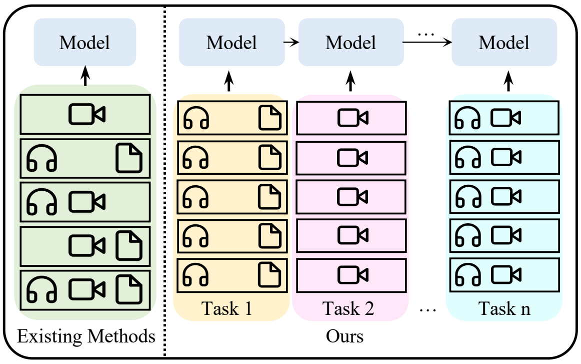

Figure 1: The difference between existing methods and ours. Existing methods train all cases of data together, which is infeasible in many real-world scenarios. In contrast, we formulate it as a continual learning problem, which is much closer to real situations.

However, in real-world dynamic environments, data often comes in a sequence where each dataset has the same modality missing (Figure 1 (right)). For example, a robot needs to utilize multiple sensors to capture human faces, gestures and speech to analyze sentiment and emotion. When the camera is broken, the system needs to make predictions without video modality during the period until the camera is repaired. During this period, all the data has the same missing modality (*i.e.* video). When the recording device is broken, the system needs to learn how to make accurate predictions without audio modality until the recording device is repaired. In such dynamic environments, the system is expected to adapt to the different missing modality cases continually. Therefore, existing methods relying on recovering missing features and aligning multimodal features will fail. Additionally, as

[*] Corresponding author

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4317–4327
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

shown in Figure 2, the sequential data in real-world applications will make these methods suffer catastrophic forgetting (McCloskey and Cohen, 1989), leading to performance degradation. To handle forgetting, an intuitive idea is to store and retrain all old data but it creates large storage overheads and potential privacy issues.

Based on the above observations, we propose the continual multimodal missing modality task to address the missing modality issues in real-world continual environments. In recent years, continual learning has made great progress, such as replay-based methods (Rolnick et al., 2019; Buzzega et al., 2020; Cha et al., 2021), regularization-based methods (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018), and architecture-based methods (Serra et al., 2018; Li et al., 2019; Ebrahimi et al., 2020). However, these methods often have many limitations. For example, replay-based methods need to store previous data, which could pose potential privacy issues. More recently, prompt-based continual methods (Wang et al., 2022a,c,b) have attracted much attention due to their simplicity and effectiveness. Most of these methods (Wang et al., 2022c,b) are unimodal and are difficult to transfer to the multimodal field. Multimodal methods (Wang et al., 2022a; Qian et al., 2023) always depend on language-image models such as CLIP (Radford et al., 2021), which makes it difficult to apply these methods to other fields where there are more modalities. Moreover, these multimodal methods focus more on exploring task interaction while ignoring modality interaction.

In this paper, we propose three types of prompts and a task interaction strategy for efficient continual multimodal missing modality task. Specifically, we propose modality-specific prompts, task-aware prompts, and task-specific prompts. Modality-specific prompts aim to instruct the model to learn intra-modality features. Task-aware prompts focus on learning inter-modality and inter-task features. Task-specific prompts help the model learn intra-task features. Moreover, we propose a contrastive task interaction strategy to grasp the relationships between tasks.

We conduct extensive experiments on three multimodal datasets: CMU-MOSI (Zadeh et al., 2016), IEMOCAP (Busso et al., 2008) and CH-SIMS (Yu et al., 2020). Our proposed method can consistently outperform baselines and state-of-the-art methods significantly in all three datasets. Besides, the number of trainable parameters only accounts for 2-3% of the parameters of the backbone network, indicating our method is parameter-efficient. We further conduct ablation experiments to validate the effectiveness of three types of prompts and contrastive task interaction strategy. The results fully demonstrate the superiority of our method. Our main contributions can be summarized as follows:

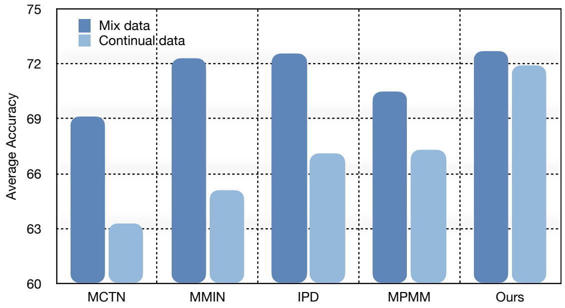

Figure 2: The performance of existing methods will degrade when applied to continual multimodal missing modality task.

- We introduce a comprehensive formulation of *continual multimodal missing modality* task.
- We propose modality-specific prompts, task-aware prompts, task-specific prompts and a contrastive task interaction strategy. They can be transferred easily to any multimodal backbones efficiently.
- We build up three benchmarks for continual multimodal missing modalities. Our proposed method outperforms all the baselines and state-of-the-art approaches significantly.

2 Related Work

Multimodal Learning with Missing Modalities. Missing modality issues pose challenges for multimodal learning (Guo et al., 2024b) and can lead to severe performance degradation. Recently, many works explore to address the missing modality issues (Ma et al., 2021; Cai et al., 2018; Du et al., 2018; Zhao et al., 2021; Lee et al., 2023; Jin et al., 2023). Some methods (Cai et al., 2018; Du et al., 2018) directly generate missing modalities using the available modalities. Pham et al. (2019) propose to align multimodal features by translating between modalities to address missing modality issues. Zhao et al. (2021) propose learning robust joint multimodal representations that can predict the representation of any missing modality given the available modalities. IPD (Jin et al., 2023)

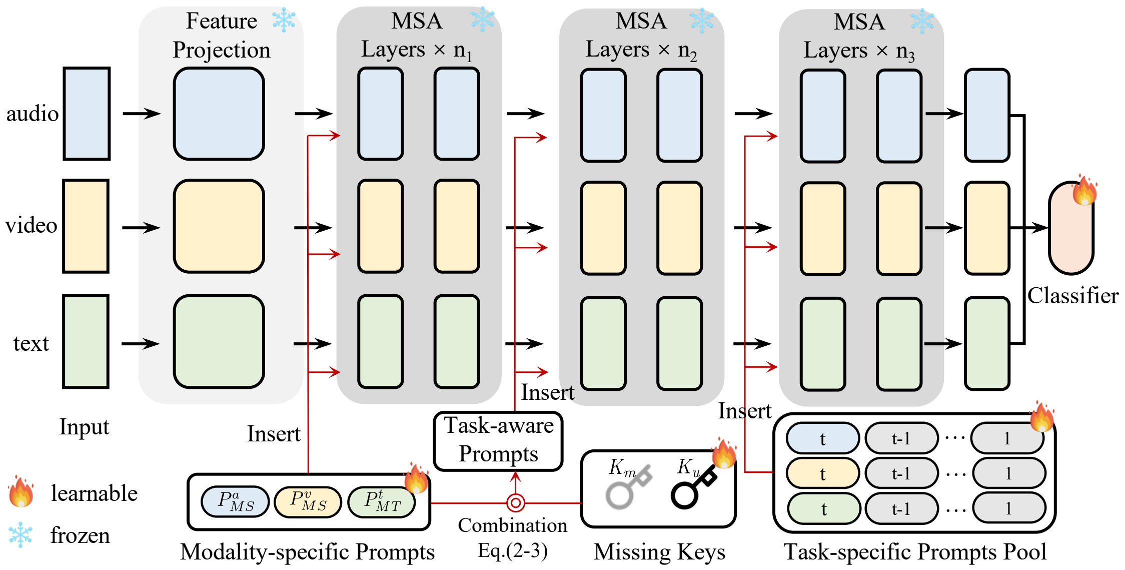

Figure 3: The overall architecture of our proposed method. After the projection layer, modality-specific prompts, task-aware prompts and task-specific prompts are attached to multi-head self-attention (MSA) layers sequentially. Task-aware prompts are generated from modality-specific prompts and missing keys using Eq.(2).

jointly learns modality-specific task prototypes. Guo et al. (2024c) propose three types of prompts to address missing modality issues in a parameter-efficient way. Guo and Jin (2025) propose to address missing modalities at test time by smoothing the distribution shifts between the complete data samples and incomplete data samples.

Continual Learning. A major challenge of continual learning is known as catastrophic forgetting (McCloskey and Cohen, 1989). Numerous methods have been exploited to address this issue which could be categorized into three main categories: (1) Regularization-based approaches (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018) address catastrophic forgetting by imposing a regularization constraint to important parameters. (2) Replay-based approaches (Rebuffi et al., 2017; Shin et al., 2017; Rolnick et al., 2019; Buzzega et al., 2020; Cha et al., 2021) store some representative samples of previous tasks in a rehearsal buffer and retrain these data to avoid forgetting. (3) Architecture-based approaches (Mallya and Lazebnik, 2018; Serra et al., 2018; Li et al., 2019; Ebrahimi et al., 2020) dynamically expand or divide the network for different tasks to mitigate forgetting. These methods often suffer from scalability issues as the number of tasks or the complexity of the model increases. Our proposed method is based on prompt learning and is a replay-free method. Moreover, our novel design of prompts can instruct the model to address complex situations compared to regularization-based methods and architecture-based methods.

Prompt Learning. Prompt learning, as one of the efficient transfer learning techniques (Hu et al., 2021; Guo et al., 2024a; Yan et al., 2025), refers to the process of designing or generating effective prompts to use a pre-trained model for different types of downstream tasks. Recent works (Wang et al., 2022a,c,b; Yan et al., 2024) apply prompt learning to the field of continual learning and have achieved good results. DualPrompt (Wang et al., 2022b) proposes G-Prompt and E-Prompt to learn task-invariant and task-specific information, but it is unimodal and can not be directly transferred to multimodal applications. Particularly, Wang et al. (2022a) propose S-Prompts which is multimodal, but this prompting method ignores the modality-level information. Moreover, S-Prompts is a CLIP-based (Radford et al., 2021) approach which is a language-image scheme and thus can not address problems which has more modalities. In contrast, our proposed method has both modality interaction and task interaction strategies and can be easily transferred to any backbones.

3 Proposed Method

3.1 Problem Formulation

In real-world dynamic environments, the new data come continually which could have different modality cases. Therefore, we can consider

Table 1: The seven different missing modality cases and their denotations.

No.	{available, missing}	denotation
1	$\{(a, v, t), ()\}$	$\boldsymbol{x} = (x^a, x^v, x^t)$
2	$\{(a, v), (t)\}$	$\boldsymbol{x} = (x^a, x^v, x^{tm})$
3	$\{(a, t), (v)\}$	$\boldsymbol{x} = (x^a, x^{vm}, x^t)$
4	$\{(v, t), (a)\}$	$\boldsymbol{x} = (x^{am}, x^v, x^t)$
5	$\{(a), (v, t)\}$	$\boldsymbol{x} = (x^a, x^{vm}, x^{tm})$
6	$\{(v), (a, t)\}$	$\boldsymbol{x} = (x^{am}, x^v, x^{tm})$
7	$\{(t), (a, v)\}$	$\boldsymbol{x} = (x^{am} x^{vm}, x^t)$

it as a domain-incremental learning task. In a common domain-incremental learning setting, training samples of different domains arrive in sequence (*i.e.* data with different missing modality cases in our task). We denote the sequential datasets as $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_T\}$, where $\mathcal{D}_t = \{(\boldsymbol{x}_i^t, y_i^t)\}_{i=1}^{N_t}$ represents the dataset for the t-th task with N_t training samples. For example, as shown in Table 1, $\mathcal{D}_4$ represents the dataset with audio modality missing. In this paper, we consider a case of $M = 3$ modalities (audio, video and text) for simplicity. Therefore, $\boldsymbol{x}_i^t$ consists of three modalities and there are a total of $2^M - 1 = 7$ different missing modality cases (shown in Table 1).

3.2 Prompt Design

Existing methods address missing modalities mainly by utilizing complicated modules to generate missing information (Zhao et al., 2021; Du et al., 2018) or aligning multimodal representations (Pham et al., 2019). Besides, existing continual methods often cause privacy issues or scalability issues. Motivated by prompt learning methods (Wang et al., 2022b; Lee et al., 2023; Guo et al., 2024c), we propose three types of prompts to address missing modality cases in a continual setting, which is simple and computationally efficient. Specifically, our proposed method contains three types of prompts: modality-specific prompts, task-aware prompts, and task-specific prompts (as shown in Figure 3).

Modality-specific Prompts. Existing prompt-based methods (Wang et al., 2022a,c,b) mainly focus on task interaction while ignoring interactions between modalities. Therefore, to model inter-modality features, we propose modality-specific prompts. We denote modality-specific prompts as $P_{MS} \in \mathbb{R}^{M \times \ell \times d}$ where ℓ and d represent the length and dimension of the prompt respectively. P_{MS} consists of P^a_{MS}, P^v_{MS} and P^t_{MS}, which represent audio, video and text modality, respectively. The modality-specific prompts are modality-specific but task-shared. We attach this kind of prompt to the following n_1 multi-head self-attention (MSA) layers after the feature projection layer, where n_1 is a hyperparameter. The process of attaching prompts to the i-th MSA layer is:

$$\begin{aligned} h_a^{(i)} &= \texttt{A-MSA}^{(\mathrm{i})}([P^a_{MS}; h_a^{(i-1)}]) \\ h_v^{(i)} &= \texttt{V-MSA}^{(\mathrm{i})}([P^v_{MS}; h_v^{(i-1)}]) \\ h_t^{(i)} &= \texttt{T-MSA}^{(\mathrm{i})}([P^t_{MS}; h_t^{(i-1)}]) \end{aligned} \tag{1}$$

where $[\cdots]$ is the concatenation operation along the sequence, h^i_m is the feature representation of modality m after the i-th MSA layer, and $\texttt{A-MSA}^{(\mathrm{i})}$, $\texttt{V-MSA}^{(\mathrm{i})}$ and $\texttt{T-MSA}^{(\mathrm{i})}$ represent the i-th audio, video and text MSA layer, respectively.

Task-aware Prompts. Given the input x, the model should be informed of the missing condition of x to address missing information. Therefore, we propose task-aware prompts to learn the inter-modality features between the missing modalities and available modalities. To generate task-aware prompts, we introduce missing keys which are a sign of whether a modality is missing or not. Specifically, we denote missing keys as $K = \{K_m, K_u\}$ where K_m represents a modality is missing while K_u represents a modality is available. $K_m, K_u \in \mathbb{R}^d$, which are also trainable parameters. Concretely, we use the following equations to generate task-aware prompts for each modality:

$$P_k = \beta_k \cdot K_m \odot P^k_{MS} + (1 - \beta_k) \cdot K_u \odot P^k_{MS} \tag{2}$$

where $\odot$ is the element-wise multiplication of the broadcasted vector and the matrix and P^k_{MS} are the modality-specific prompts. $k \in \{a, v, t\}$, $\beta_k \in \{0, 1\}$ is a sign function to denote whether the modality k is missing. $\beta_k = 0$ represents the modality k is missing and $\beta_k = 1$ represents the modality k is available. It is worth noting that modality-specific prompts and missing keys are both task-agnostic, but their combinations are task-dependent. This design can not only reduce the number of trainable parameters but also connect the intra- and inter-modality features. Then, we can obtain the task-aware prompts P_{TA} as follows:

$$P_{TA} = P_a + P_v + P_t \tag{3}$$

After we obtain the task-aware prompts P_{TA}, we attach them to the next n_2 MSA layers following the first n_1 layers which are attached with modality-specific prompts. The prompts are attached to the i-th MSA layer as follows:

$$\begin{aligned} h_a^{(i)} &= \mathsf{A\text{-}MSA}^{(i)}([P_{TA}; h_a^{(i-1)}]) \\ h_v^{(i)} &= \mathsf{V\text{-}MSA}^{(i)}([P_{TA}; h_v^{(i-1)}]) \\ h_t^{(i)} &= \mathsf{T\text{-}MSA}^{(i)}([P_{TA}; h_t^{(i-1)}]) \end{aligned} \tag{4}$$

Comparing Eq.(1) and Eq.(4), it is easy to discover that the main difference is that different modality has different prompts in Eq.(1) but all modalities have the same prompts in Eq.(4). That is also the main difference between these two types of prompts. Modality-specific prompts are modality-specific but task-agnostic and task-aware prompts are task-dependent but modality-shared. This indicates that modality-specific prompts focus more on learning intra-modality information and task-aware prompts more on inter-modality information.

Task-specific Prompts. Although task-aware prompts are different in different tasks, they are generated from modality-specific prompts and missing keys which are shared by all tasks. The main role of task-aware prompts is to help the model learn inter-modality information and inform the model of the missing modality condition. Therefore, modality-specific prompts and task-aware prompts are not able to learn task-specific information to address catastrophic forgetting. Based on this observation, we propose task-specific prompts $P_{TS} = \{P_{TS}^{(1)}, P_{TS}^{(2)}, \cdots, P_{TS}^{(T)}\}$ to instruct the model for a specific task and address catastrophic forgetting. Specifically, for every task t, we have task-specific prompts $P_{TS}^{(t)} \in \mathbb{R}^{M \times \ell \times d}$. Moreover, $P_{TS}^{(t)} = \{P_{TS_a}^{(t)}, P_{TS_v}^{(t)}, P_{TS_t}^{(t)}\}$. We attach the prompts the same as before:

$$\begin{aligned} h_a^{(i)} &= \mathsf{A\text{-}MSA}^{(i)}([P_{TS_a}^{(t)}; h_a^{(i-1)}]) \\ h_v^{(i)} &= \mathsf{V\text{-}MSA}^{(i)}([P_{TS_v}^{(t)}; h_v^{(i-1)}]) \\ h_t^{(i)} &= \mathsf{T\text{-}MSA}^{(i)}([P_{TS_t}^{(t)}; h_t^{(i-1)}]) \end{aligned} \tag{5}$$

As before, we attach the prompts to n_3 MSA layers after the $n_1 + n_2$ MSA layers which are attached with modality-specific prompts and task-aware prompts.

3.3 Task Interaction Strategy

Unlike many common domain-incremental learning tasks where there are no evident relationships between the domains, continual multimodal missing modality task has some implicit relationships between different domains. For example, text and audio are always highly relevant because they are both high-semantic information. Different from text and audio, videos contain facial expressions or gestures which are low-semantic information. Therefore, in the representation space, text often has features very similar to that of audio, but much different from that of video.

Based on this observation, we propose to consider audio prompts and text prompts as very similar instances and make them close together in the representation space while making audio and text prompts far from video prompts. Specifically, we adopt a contrastive scheme for task-aware prompts. We consider task-aware prompts of task $\boldsymbol{x} = (x^a, x^v, x^{tm})$ and task $\boldsymbol{x} = (x^{am}, x^v, x^t)$, task $\boldsymbol{x} = (x^a, x^{vm}, x^{tm})$ and task $\boldsymbol{x} = (x^{am}, x^{vm}, x^t)$ as positive pairs and others (except $\boldsymbol{x} = (x^a, x^v, x^t)$) as negative pairs. By doing this, task-aware prompts can learn the correlation between different missing modality cases (*i.e.* different tasks), thus strengthening the inter-task relationship.

We consider a modified *NT-Xent* loss (Chen et al., 2020) as our loss function. Let $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / ||\boldsymbol{u}|| ||\boldsymbol{v}||$ denote the dot product between ℓ_2 normalized $\boldsymbol{u}$ and $\boldsymbol{v}$ (*i.e.* cosine similarity). The loss function for a positive example (i, j) is:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{t=1}^{T} \mathbb{1}_{[t \neq i,j]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_t)/\tau)} \tag{6}$$

where τ is a temperature parameter and $\mathbb{1}_{[t \neq i,j]} \in \{0, 1\}$ is a sign function evaluating to 1 if $t \neq i, j$. We take the average value of task-aware prompts along the sequence length dimension as $\boldsymbol{z}$. As shown in Table 1, we denote the index of task-aware prompts of task $\boldsymbol{x} = (x^a, x^v, x^{tm})$, task $\boldsymbol{x} = (x^{am}, x^v, x^t)$, task $\boldsymbol{x} = (x^a, x^{vm}, x^{tm})$, and task $\boldsymbol{x} = (x^{am}, x^{vm}, x^t)$ as $2, 4, 5, 7$, respectively. Therefore, we can define our contrastive loss as:

$$\mathcal{L}_{con} = \ell_{2,4} + \lambda_2 \ell_{5,7} \tag{7}$$

where λ_2 is a trade-off between the two losses.

3.4 Overall Objective

We combine the task loss with contrastive loss as:

$$\mathcal{L} = \mathcal{L}_{task}(\hat{y}(\boldsymbol{x}), y) + \lambda_1 \mathcal{L}_{con} \tag{8}$$

where $\hat{y}(\boldsymbol{x})$ is the network prediction, y is the label, λ_1 is a hyperparameter to balance the two losses, and $\mathcal{L}_{task}$ is the task-specific loss, *e.g.* cross-entropy loss, L2 loss.

Table 2: Quantitative results on CMU-MOSI, IEMOCAP and CH-SIMS datasets. **Bold**: best exemplar-free results. Underline: second best exemplar-free results. * denotes best replay-based results. Lowerbound: training the backbone without any prompts on the continual datasets. Upperbound: supervised finetuning on the i.i.d data of all tasks. Upperbound (ours): supervised finetuning with modality-specific prompts and task-aware prompts on the i.i.d data of all tasks. AA: average accuracy, FM: forgetting measure.

Method	**Buffer size**	**CMU-MOSI**		**Buffer size**	**IEMOCAP**		**Buffer size**	**CH-SIMS**	
		AA (↑)	FM (↓)		AA (↑)	FM (↓)		AA (↑)	FM (↓)
iCaRL (Rebuffi et al., 2017)	250	64.12	3.49	500	54.63	6.11	250	63.79	3.17
A-GEM (Chaudhry et al., 2019a)		63.18	4.10		52.97	7.89		62.01	4.27
ER (Chaudhry et al., 2019b)		65.78	3.44		57.14	5.09		65.85	2.76
DER++ (Buzzega et al., 2020)		64.62	2.74		54.87	4.50		63.51	2.96
iCaRL (Rebuffi et al., 2017)	500	66.81	2.01	1000	56.46	2.39	500	65.97	1.84
A-GEM (Chaudhry et al., 2019a)		65.12	2.88		54.07	5.78		65.18	3.01
ER (Chaudhry et al., 2019b)		68.91*	1.12		58.89*	2.98		68.46*	0.97*
DER++ (Buzzega et al., 2020)		67.02	0.69*		57.56	2.40*		66.18	1.02
EWC (Kirkpatrick et al., 2017)	0	66.44	1.75	0	58.96	2.12	0	65.11	2.04
LwF (Li and Hoiem, 2017)		64.56	2.97		54.95	4.87		63.70	3.09
L2P (Wang et al., 2022c)		63.79	2.67		55.68	4.73		63.61	2.51
DualPrompt (Wang et al., 2022b)		67.23	0.73		58.15	1.29		68.73	0.89
S-Prompts (Wang et al., 2022a)		64.83	3.57		54.30	5.09		64.96	2.94
MMIM (Han et al., 2021)	-	64.25	5.31	-	52.38	9.15	-	61.37	6.96
MISA (Hazarika et al., 2020)		61.63	6.75		49.33	10.51		59.12	7.01
UniMSE (Hu et al., 2022)		64.97	5.26		52.89	9.23		53.46	6.21
MCTN (Pham et al., 2019)	-	63.35	4.17	-	56.13	5.35	-	63.11	3.94
MMIN (Zhao et al., 2021)		65.31	3.92		56.41	4.36		64.85	3.11
IPD (Jin et al., 2023)		67.13	1.84		57.63	2.91		67.16	1.41
MPLMM (Guo et al., 2024c)		70.35	2.18		60.32	2.90		68.24	2.77
Ours	-	**71.87**	**-0.15**	-	**62.24**	**0.08**	-	**71.11**	**0.04**
Lowerbound		62.34	6.18		51.15	10.31		61.18	6.87
Upperbound		71.19	-		61.74	-		70.08	-
Upperbound (Ours)		73.20	-		64.22	-		71.98	-

4 Experiments

4.1 Datasets and Evaluation Metrics

We validate our methods on CMU-MOSI, IEMOCAP and CH-SIMS.

CMU-MOSI (Zadeh et al., 2016) is a popular dataset for multimodal (audio, text and video) sentiment analysis, comprising 93 English YouTube videos which are carefully selected and divided into 2,199 segments. Each segment is manually annotated with a sentiment score ranging from strongly negative to strongly positive (-3 to +3).

IEMOCAP (Busso et al., 2008) contains recorded videos from ten actors in five dyadic conversation sessions. There are different types of emotions (happiness, anger, sadness, frustration and neutral state). In our task, four emotions (happiness, anger, sadness and neutral state) are selected for classfication.

CH-SIMS (Yu et al., 2020) is a Chinese multimodal sentiment analysis dataset. It contains 2,281 refined video segments in the wild annotated with a sentiment score ranging from strongly negative to strongly positive (-1 to 1). The dataset covers a total number of 474 distinct speakers.

For evaluation, we use Average accuracy (AA) and Forgetting measure (FM). AA is the average accuracy of all tasks and calculated as $AA = \frac{1}{n}\sum_{i=1}^{n} a_{i,n}$ where $a_{i,n}$ is the accuracy on task i after training the model on task n. FM measures the performance degradation and is calculated as $FM = \frac{1}{n-1}\sum_{i=1}^{n-1} \max_{j\in[i,n-1]}(a_{i,j} - a_{i,n})$.

4.2 Baselines

Continual methods. They include non-prompting methods: iCaRL (Rebuffi et al., 2017), EWC (Kirkpatrick et al., 2017), LwF (Li and Hoiem, 2017), A-GEM (Chaudhry et al., 2019a), ER (Chaudhry et al., 2019b), DER++ (Buzzega et al., 2020), and prompting methods: L2P (Wang et al., 2022c), DualPrompt (Wang et al., 2022b), S-Prompts (Wang et al., 2022a). For replay-based methods iCaRL, A-GEM, ER, DER++, we use two different replay buffer sizes (250, 500 for CMU-MOSI and CH-SIMS and 500, 1000 for IEMOCAP).

Robust Multimodal Methods. Besides, we compare our method with some state-of-the-art multimodal backbones: MISA (Hazarika et al., 2020), MMIM (Han et al., 2021), UniMSE (Hu et al., 2022). We replace the missing modalities with zero vectors.

Missing modality methods. Moreover, we compare our method with a series of missing modality approaches: MCTN (Pham et al., 2019), MMIN (Zhao et al., 2021), IPD (Jin et al., 2023), MPLMM (Guo et al., 2024c).

4.3 Implementation Details

For fair comparison, we use the multimodal transformer as backbone for continual learning methods. For our proposed method, the dimension d of all the prompts is set to 30 and the length ℓ is set to 16 by default. We set $n_1 = 2$, $n_2 = 3$ and $n_3 = 3$. We use L1 loss for CMU-MOSI and CH-SIMS and cross-entropy loss for IEMOCAP. After hyperparameter searching, we set $\lambda_1 = 0.1$ and $\lambda_2 = 1$. In all experiments, we use Adam optimizer with a batch size of 64. For other hyperparameters, we follow the original paper of comparing methods. We train all the models for 30 epochs with a learning rate of 1×10^{-3}.

For non-prompting methods iCaRL, EWC, LwF, A-GEM, ER, DER++, we do not freeze the backbone. For prompt-based methods L2P, DualPrompt, S-Prompts and MPLMM, we freeze the pre-trained backbone and only finetune the learnable prompts.

4.4 Main Results

Table 2 presents the performance of all methods on CMU-MOSI, IEMOCAP and CH-SIMS datasets.

Comparison with continual learning methods. Compared with replay-based methods which could lead to privacy issues, our method does not use any buffered data and still can achieve better performance than those with a memory buffer. Compared with exemplar-free continual methods, our method achieves better average results and forgetting measure, indicating the effectiveness of our proposed prompts which promote the model to learn intra-modality, inter-modality and inter-task information.

Comparison with multimodal and missing modality methods. Besides, we compare our methods with multimodal and missing modality approaches. The results reveal that multimodal methods all have low average accuracy and high forgetting measure, which indicates that they are not able to deal with missing modality issues and catastrophic forgetting. In comparison, missing modality approaches can achieve comparable or even higher average accuracy than those continual methods due to modules or strategies that are designed to address missing modalities. However, compared to continual methods, these methods often have higher forgetting measure, indicating that they fail to address the catastrophic forgetting. In contrast, our method can not only address the missing modality issue but also deal with catastrophic forgetting in the dynamic environment.

Table 3: An ablation study of three different types of prompts on CMU-MOSI.

P_{MS}	P_{TA}	P_{TS}	AA (↑)	FM (↓)
			62.34	6.18
✓			64.07	4.01
	✓		66.21	3.24
		✓	68.01	1.27
✓	✓		69.16	2.08
✓		✓	70.34	0.74
	✓	✓	70.51	0.31
✓	✓	✓	**70.91**	**0.13**

Performance of our method. We get a negative forgetting measure on CMU-MOSI, which indicates that in the process of learning new tasks, the model performs even better on previous tasks. This demonstrates the effectiveness of our novel design of prompts and task interaction strategy, which enables the model to learn better intra-modality and intra-task relationships, thus making it perform better on previous tasks without forgetting.

Moreover, our proposed method outperforms upperbound slightly which is trained on the i.i.d data of all tasks. This fully demonstrates that our method improves inter-modality communication. Comparing the upperbound and the upperbound using our designed prompts, we could also discover our method makes it easier for the model to learn intra-modality and inter-modality information.

Efficiency of our method. It is worth noting that the number of trainable parameters of our method only accounts for about 2-3% of the parameters of the backbone network. With such few parameters, our method can achieve better results than other baseline methods, which indicates that our method is parameter-efficient and effective.

4.5 Ablation Study

Effectiveness of three types of prompts. In Table 3, we show quantitative results of the benefits of three types of prompts. It is easy to find

Table 4: An ablation study on the benefit of task interaction strategy on CMU-MOSI.

Method	AA (↑)	FM (↓)
w/o. task interaction	70.91	0.13
w/o. $\ell_{2,4}$, $\lambda_1 = 0.1$	71.14	0.09
w/o. $\ell_{5,7}$, $\lambda_1 = 0.1$	71.42	0.06
$\lambda_1 = 0.2, \lambda_2 = 1$	71.52	0.04
$\lambda_1 = 0.1, \lambda_2 = 2$	71.69	0.01
$\lambda_1 = 0.1, \lambda_2 = 0.5$	71.58	-0.04
$\lambda_1 = 0.1, \lambda_2 = 1$	**71.87**	**-0.15**

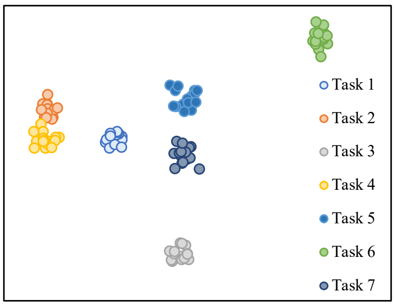

Figure 4: t-SNE visualization of task-aware prompts on the CMU-MOSI dataset. Each point represents a prompt vector. Tasks 1-7 are shown in Table 1.

that task-specific prompts contribute most to addressing catastrophic forgetting. As shown in the table, the forgetting measure of the model with only task-specific prompts P_{TS} is lower than that of the model with modality-specific prompts P_{MS} and task-aware prompts P_{TA}. This indicates that task-specific prompts help the model learn intra-task information, which plays a very important role in dealing with forgetting. Besides, modality-specific prompts and task-aware prompts help a lot in improving the model's average accuracy. Modality-specific prompts help the model learn intra-modality information and task-aware prompts help the model learn inter-modality and inter-task information. The combination of three types of prompts further enhances the performance of the model, which fully convinces us of the effectiveness of our proposed prompts.

Moreover, we visualize task-aware prompts using t-SNE in Figure 4. We can observe that points of Task 2 and Task 3, Task 5 and Task 7 are very close to each other. This indicates the effectiveness of our task interaction strategy, which helps the model learn inter-task relationships. Besides, task-aware prompts of different tasks are well-separated, which demonstrates that these prompts help the model learn task-dependent knowledge.

Table 5: An ablation study of the sequence of attaching these three types of prompts on CMU-MOSI. $A \rightarrow B \rightarrow C$ represents that we attach A prompts at the first n_1 MSA layers, B prompts at the following n_2 MSA layers, and C prompts at the next n_3 MSA layers. Here, we set $n_1 = 2$, $n_2 = 3$, $n_3 = 3$. **Bold**: best results.

Prompt Sequence	AA (↑)	FM (↓)
$P_{MS} \rightarrow P_{TA} \rightarrow P_{TS}$	**71.87**	**-0.15**
$P_{MS} \rightarrow P_{TS} \rightarrow P_{TA}$	70.96	-0.03
$P_{TA} \rightarrow P_{MS} \rightarrow P_{TS}$	71.23	0.06
$P_{TA} \rightarrow P_{TS} \rightarrow P_{MS}$	70.57	0.04
$P_{TS} \rightarrow P_{MS} \rightarrow P_{TA}$	70.00	0.10
$P_{TS} \rightarrow P_{TA} \rightarrow P_{MS}$	69.98	0.28

Table 6: An ablation study of the specific positions of prompts on the CMU-MOSI dataset. Our backbone has ten MSA layers in total. **Bold**: best results.

P_{MS}	P_{TA}	P_{TS}	AA (↑)	FM (↓)
$[1, 2]$	$[3, 4, 5]$	$[6, 7, 8, 9]$	70.06	0.23
$[1, 2]$	$[3, 4, 5]$	$[6, 7, 8]$	**71.87**	**-0.15**
$[1, 2]$	$[3, 4, 5]$	$[6, 7]$	71.10	0.07
$[1, 2, 3]$	$[4, 5]$	$[6, 7, 8]$	71.42	-0.09

Effectiveness of task interaction strategy. In Section 3.3, we introduce a task interaction strategy. To demonstrate the effectiveness of our proposed task interaction strategy, we present our ablation results in Table 4. We find that the model with two loss terms performs much better on average accuracy and forgetting measure than that without the loss terms. Besides, the model without task interaction strategy performs worse than the upperbound method shown in Table 2. This indicates that our proposed task interaction strategy acts as a bridge between tasks and helps the model learn the inter-task information, thus outperforming the upperbound method. In the fourth to sixth rows of the table, we explore the impact of the trade-off terms λ_1 and λ_2 on the performance of the model. The results reveal that the performance of the model is not sensitive to the value of λ_1 and λ_2.

Exploration of where to attach prompts. We first conduct a series of experiments to explore the sequence of three types of prompts and present our results in Table 5. We can find that the model with task-aware prompts P_{TA} in front of task-specific prompts P_{TS} always outperforms the model with P_{TA} behind P_{TS}. This indicates that compared to task-specific prompts, task-aware prompts learn low-level features, serving as a guide-

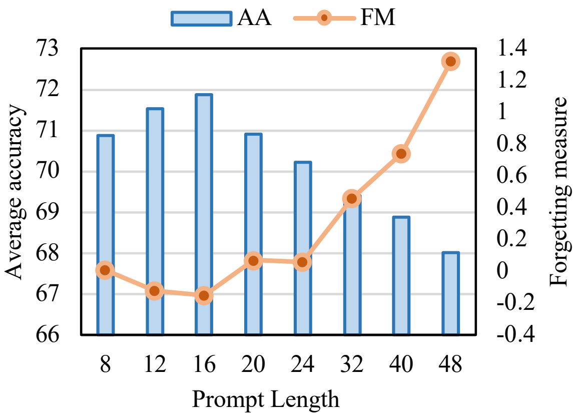

Figure 5: Quantitative results on the CMU-MOSI dataset with different prompt lengths ℓ.

line to task-specific prompts and helping task-specific prompts learn better intra-task information. Besides, modality-specific prompts instruct the model to learn intra-modality information which are low-level features at early stages. Therefore, modality-specific prompts should be placed in front of the other two types of prompts.

Furthermore, we explore the specific positions of these prompts and present our results in Table 6. Comparing the results in the first row and the second row in the table, we find that it is not opportune to attach prompts at the back layers of the network. The highest performance demonstrates the effectiveness of our design of prompts.

Impact of the length of prompts. To study the impact of prompt length on our model, we train our model on CMU-MOSI with eight different prompt lengths and present results in Figure 5. Intuitively, the longer the prompt length, the better the performance of the model. However, as the results are shown in the figure, we find that with the increasing length ℓ, the performance first improves and then declines with the peak performance at $\ell = 16$. This suggests that our proposed method can achieve great results with a relatively small number of parameters.

5 Conclusion

In this paper, we introduce the task of *continual multimodal missing modality* to tackle the challenges posed by missing modalities in dynamic environments. We propose a novel and efficient prompt design consisting of three distinct types of prompts, complemented by a contrastive task interaction strategy aimed at mitigating catastrophic forgetting in the multimodal domain. Our approach facilitates effective learning of intra-modality, inter-modality, intra-task, and inter-task features, enhancing the model's adaptability. Extensive experiments and ablation studies validate the robustness and efficacy of our proposed method. Given that cases of missing modalities frequently arise during data collection in real-world scenarios, we believe our approach represents a significant step towards practical applications in multimodal fields facing ongoing missing modality challenges.

Limitations

In our approach, the number of task-specific prompts is the same as the number of tasks. However, the number of tasks increases exponentially as the number of modalities increases. Therefore, when there are many modalities, it would cost large computational resources. Beyond this work, we believe some promising future works would solve this problem.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930.

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1158–1166, New York, NY, USA. Association for Computing Machinery.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. 2021. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 9516–9525.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019a. Efficient lifelong learning with a-gem. In *ICLR*.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019b. On tiny episodic memories in continual learning.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*.

Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM international conference on Multimedia*. ACM.

Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. 2020. Adversarial continual learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 386–402. Springer.

Zirun Guo, Xize Cheng, Yangyang Wu, and Tao Jin. 2024a. A wander through the multimodal landscape: Efficient transfer learning via low-rank sequence multimodal adapter. *arXiv preprint arXiv:2412.08979*.

Zirun Guo and Tao Jin. 2025. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *The Thirteenth International Conference on Learning Representations*.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. 2024b. Classifier-guided gradient modulation for enhanced multimodal learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024c. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, Bangkok, Thailand. Association for Computational Linguistics.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tao Jin, Xize Cheng, Linjun Li, Wang Lin, Ye Wang, and Zhou Zhao. 2023. Rethinking missing modality learning from a decoding perspective. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 4431–4439, New York, NY, USA. Association for Computing Machinery.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, pages 3925–3934. PMLR.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Wang Lin, Jingyuan Chen, Jiaxin Shi, Zirun Guo, Yichen Zhu, Zehan Wang, Tao Jin, Zhou Zhao, Fei Wu, Shuicheng YAN, and Hanwang Zhang. 2024. Action imitation in common action space for customized action image synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310.

Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. 2023. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2953–2962.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. 2022a. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. 2022b. Dual-prompt: Complementary prompting for rehearsal-free continual learning. *European Conference on Computer Vision*.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022c. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149.

Weicai Yan, Wang Lin, Zirun Guo, Ye Wang, Fangming Feng, Xiaoda Yang, Zehan Wang, and Tao Jin. 2025. Diff-prompt: Diffusion-driven prompt generator with mask supervision. In *The Thirteenth International Conference on Learning Representations*.

Weicai Yan, Ye Wang, Wang Lin, Zirun Guo, Zhou Zhao, and Tao Jin. 2024. Low-rank prompt interaction for continual vision-language retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8257–8266.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, Online. Association for Computational Linguistics.