MaCSC: Towards Multimodal-augmented Pre-trained Language Models via Conceptual Prototypes and Self-balancing Calibration

Xianwei Zhuang, Zhichang Wang, Xuxin Cheng, Yuxin Xie,
Liming Liang, Yuexian Zou*
School of ECE, Peking University, China
{xwzhuang,wzcc,chengxx,yuxinxie,limingliang}@stu.pku.edu.cn,
zouyx@pku.edu.cn

Abstract

Pre-trained language models (PLMs) that rely solely on textual data may exhibit limitations in multimodal semantics comprehension. Existing solutions attempt to alleviate this issue by incorporating explicit image retrieval or generation techniques. However, these methods: (1) focus exclusively on the static image modality; (2) inevitably encounter modality gaps and noise; (3) indiscriminately treat all modalities. In this paper, we propose a novel multimodal-augmented framework termed MaCSC, which can infuse multimodal semantics into PLMs and facilitate a self-balancing calibration of information allocation. Specifically, MaCSC obtains modal-specific conceptual prototypes from contrastive pre-training models (e.g., CLIP), and aggregates the intra- and inter-modal semantics of the conceptual prototype to enhance PLMs. In addition, we utilize a novel self-balancing contrastive loss to achieve multi-scale self-balancing calibration of multimodal information during fine-tuning PLMs. Experimental results show that MaCSC consistently improves the performance of PLMs across various architectures and scales, and outperforms competitive baselines on multiple NLP tasks.

1 Introduction

Large-scale Pretrained Language Models (PLMs), (e.g., BERT (Devlin et al., 2019),T5 (Raffel et al., 2020), GPT (Brown et al., 2020), etc.) have achieved great success on various Natural Language Understanding (NLU) tasks. Most PLMs undergo training on textual data alone without incorporating information from other modalities. Those approaches contrast with human learning (Bender and Koller, 2020; Tan and Bansal, 2020a), which typically involves multiple modalities beyond just text (e.g., image, audio, video, etc). Textual data may not fully capture the general characteristics of entities (Liu et al., 2022), and word frequency may not consistently reflect real-world likelihoods (Gordon and Durme, 2013). Consequently, these language learners may lack real-world knowledge (Zhang et al., 2022) and could suffer from human reporting bias (Yang et al., 2022).

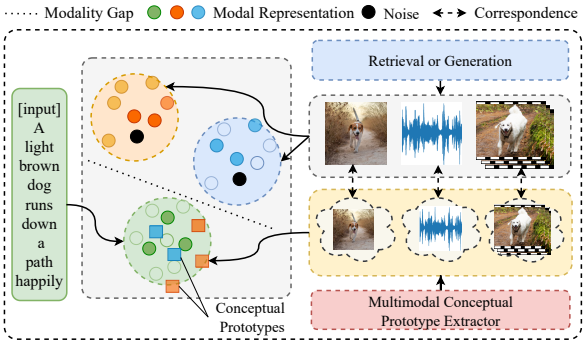


Figure 1: The approach for augmenting PLMs by retrieving and generating relevant multimodal data inevitably encounters challenges associated with (a) modal gaps and (b) the presence of noisy data. Our methodology employs conceptual prototype extractors to effectively mitigate the issues of modality gaps and noisy data, and incorporate multimodal knowledge into PLMs.

To address this issue, current research primarily focuses on augmenting PLMs by integrating visual information from either retrieved or synthesized images in NLP tasks (Tan and Bansal, 2020a; Yang et al., 2022). We classify these visually-augmented approaches for PLMs into three distinct categories: visually-augmented pre-training (Tan and Bansal, 2020a; Wang et al., 2022), visually-augmented fine-tuning (Lu et al., 2022; Guo et al., 2023), and training-free zero-shot visual enhancement (Yang et al., 2022). While these works represent notable advancements, we discover several limitations: (1) They focus exclusively on static image modality while neglecting other modal types like speech and video, which offer dynamic temporal information. (2) They require time-consuming retrieval or explicit image generation and inevitably encounter issues with modality gaps and data noise. As il-

*Corresponding author.

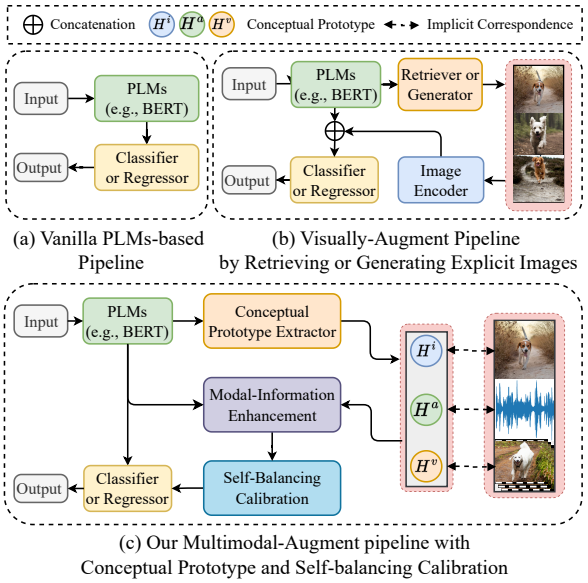
Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8077–8090

lustrated in Figure 2, a significant semantic gap exists between different modalities in the representation space (Liang et al., 2022). This modality gap can adversely impact the integration of multimodal information. Moreover, the processes of image retrieval or generation are prone to introducing noise, which can degrade the performance of PLMs. (3) They treat all modalities of information equally. Recent studies (Dai et al., 2022; Guo et al., 2023) indicate that information from other modalities is not always beneficial for NLP tasks. The contributions of various modal types vary across specific NLP tasks. Hence, indiscriminately incorporating semantic information from multiple modalities can be suboptimal, potentially diminishing the efficacy of PLMs.

In this paper, we propose a novel **m**ultimodal-**a**ugmented framework with **c**onceptual prototypes enhancement and **s**elf-balancing **c**alibration (MaCSC). MaCSC tackles the above issues through innovative strategies in the following two aspects:

(1) **MaCSC leverages conceptual prototypes as multimodal proxies to infuse multimodal knowledge into PLMs.** MaCSC does not explicitly introduce multimodal semantics by retrieving or generating corresponding modal data. MaCSC initially obtains text semantics aligned with other modalities through large-scale multimodal contrastive pre-training models (MC-PTMs) (e.g., CLIP (Radford et al., 2021), AudioCLIP (Guzhov et al., 2021) and CLIP-ViP (Xue et al., 2022), etc.). These text semantics intrinsically have an implicit correspondence with real-world modal data, which we refer to as conceptual prototypes. Conceptual prototypes contain modal-specific representation encodings and can serve as implicit proxies. This approach alleviates representation degradation caused by modality gaps and noise. We analyze the theoretical and practical effectiveness of this prototype proxies in Section 2.2 and Section 3.2.

(2) **MaCSC implements a multi-scale self-balancing calibration for the allocation of multimodal information.** The allocation weights assigned to various modal types vary across specific NLP tasks. MaCSC first implements fine-grained and coarse-grained allocation estimation and achieves balanced information injection for PLMs through a proposed self-balancing contrastive loss. We emphasize fine-tuning PLMs through the calibration of cross-modal contrastive learning, which gives a disregarded but significant insight into balancing the allocation of multimodal information.



(a) Vanilla PLMs-based Pipeline

(b) Visually-Augment Pipeline by Retrieving or Generating Explicit Images

(c) Our Multimodal-Augment pipeline with Conceptual Prototype and Self-balancing Calibration

Figure 2: We visually demonstrate a comparison between the learning architectures of various language models. (a) acquires knowledge by fine-tuning vanilla PLMs, while (b) augments PLMs by explicitly retrieving or generating images. (c) recalls modality-specific conceptual prototypes related to sentences and achieves balanced calibration of modal information allocation, thereby implicitly enhancing language comprehension.

Figure 2 shows the main differences between MaCSC and other methods. Extensive experiments conducted on ten datasets across three NLP tasks (e.g., natural language understanding, Q&A, and text generation tasks) demonstrate the effectiveness and universality of MaCSC. Our contributions are: (1) We propose a novel multimodal-augmented framework called MaCSC, which integrates multimodal semantics into PLMs through conceptual prototypes and enhancement modules. (2) MaCSC utilizes a proposed novel self-balancing calibration strategy to balance multimodal information. (3) Numerous experiments on multiple NLP tasks demonstrate the significant performance of MaCSC.

2 Methods

In this section, we first introduce the task setting and overview of our method in Section 2.1, and then describe our proposed conceptual prototype extraction approach in Section 2.2 and cross-modal semantic enhancement in Section 2.3. Subsequently, we describe proposed allocation estimation strategies and self-balancing contrastive learning in Section 2.4 and Section 2.5, respectively.

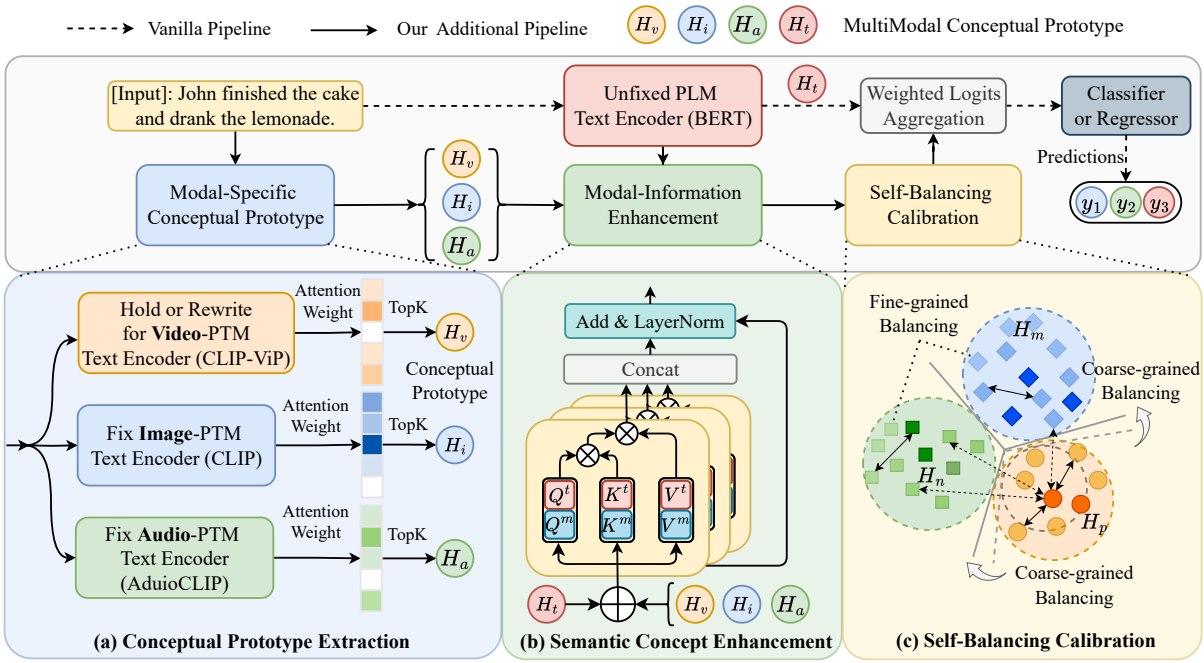


Figure 3: The illustration of the proposed framework, consisting of (a) the conceptual prototype extraction module designed to obtain implicit multimodal semantic information, (b) the multimodal semantic concept enhancement aimed at fusing cross-modal features, and (c) the multi-scale self-balancing calibration focuses on balancing modal supervision information.

2.1 Task Setting and Overview

For traditional NLP tasks, PLMs can be fine-tuned on a specified dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ to perform classification or regression, where, x_i and y_i denote i-th text data and corresponding label and N represents the total number of training samples.

In our research, we concentrate on developing an efficient approach to incorporate multi-modal knowledge into PLMs during fine-tuning. We define the set of modalities involved in this work as $\mathcal{M} = \{\mathbf{t}, \mathbf{i}, \mathbf{a}, \mathbf{v}\}$, where, $\mathbf{t}, \mathbf{i}, \mathbf{a}, \mathbf{v}$ represent four modalities: text, image, audio, and video, respectively. Our proposed method is illustrated in Figure 3, where we first generate conceptual prototypes through different MC-PTMs. Subsequently, we adopt our proposed cross-modal enhancement mechanism to obtain semantic-enhanced representations. Finally, we employ the proposed multi-scale self-balancing contrastive learning to calibrate multimodal information.

2.2 Conceptual Prototype Extraction

In our methodology, we define modal-specific conceptual prototypes as semantic units that infuse multimodal information into PLMs. Motivated by Radford et al. (2021), we employ MC-PTMs to facilitate implicit conceptual prototype extraction. We compute the average attention scores between each token and the `[EOS]` token on the self-attention layer of CLIP-based text encoder. Subsequently, we adopt the Top-K strategy to select the features of the top-ranked tokens as the modal-specific conceptual prototype $\mathbf{H}_i^m \in \mathbb{R}^{k \times d}$:

$$\mathbf{H}_i^m = \text{Top-K}(\text{TextEncoder}_m(x_i)), \quad (1)$$

where, m represents a specific modality other than text, *i.e.*, $m \in \mathcal{M} \backslash \{t\}$, k denotes the number of tokens selected, d is the dimension of features, and TextEncoder_m denotes a text encoder on a MC-PTMs trained on modality m and text-paired data. In this work, we utilize the text encoders of CLIP (Radford et al., 2021), AudioCLIP (Guzhov et al., 2021) and CLIP-ViP (Xue et al., 2022) to obtain conceptual prototypes of image, audio, and video modalities, respectively. We further provide a theoretical analysis in Appendix A.1 about conceptual prototypes are good modal proxies.

2.3 Semantic Concept Enhancement

Although conceptual prototypes represent multimodal knowledge of input sentences, they are independent of the semantics of the text acquired by PLMs. To aggregate and strengthen the representation of conceptual prototypes, we propose a semantic concept enhancement module to model

both the inter-modality and intra-modality relationships in a unified model.

We first employ a PLM to obtain the contextualized word representations $\mathbf{H}_i^t \in \mathbb{R}^{l\times d}$ of the inputs, where l denotes the length of the sequence. As shown in Figure 3, the semantic concept enhancement module takes the stacked features of conceptual prototypes and textual representations $\mathbf{E}_i = \begin{pmatrix} \mathbf{H}_i^t \\ \mathbf{H}_i^m \end{pmatrix} \in \mathbb{R}^{(k+l)\times d}$ as the input, where, m represents the modality type and $m \in \mathcal{M}\backslash\{t\}$.

Following Wei et al. (2020), the query, key, value, and cross-modal attention for the fragments are formed with the following equations:

$$\begin{aligned}
\mathbf{K_E} &= \begin{pmatrix} \mathbf{H}_i^t\mathbf{W}^K \\ \mathbf{H}_i^m\mathbf{W}^K \end{pmatrix} = \begin{pmatrix} \mathbf{K}^t \\ \mathbf{K}^m \end{pmatrix}, \\
\mathbf{Q_E} &= \begin{pmatrix} \mathbf{H}_i^t\mathbf{W}^Q \\ \mathbf{H}_i^m\mathbf{W}^Q \end{pmatrix} = \begin{pmatrix} \mathbf{Q}^t \\ \mathbf{Q}^m \end{pmatrix}, \\
\mathbf{V_E} &= \begin{pmatrix} \mathbf{H}_i^t\mathbf{W}^V \\ \mathbf{H}_i^m\mathbf{W}^V \end{pmatrix} = \begin{pmatrix} \mathbf{V}^t \\ \mathbf{V}^m \end{pmatrix}, \\
\hat{\mathbf{E}}_i &= \begin{pmatrix} \hat{\mathbf{H}}_i^t \\ \hat{\mathbf{H}}_i^m \end{pmatrix} = \mathrm{softmax}(\frac{\mathbf{Q_E}\mathbf{K_E}^\top}{\sqrt{d}})\mathbf{V_E},
\end{aligned} \tag{2}$$

where, $\hat{\mathbf{H}}_i^t \in \mathbb{R}^{l\times d}$ and $\hat{\mathbf{H}}_i^m \in \mathbb{R}^{k\times d}$ are the enhanced semantic features.

To simplify our derivation and enhance its comprehensibility, we have omitted the softmax and scaling functions in the aforementioned equation. This exclusion does not detract from the fundamental concept of our attention mechanism. The expanded form is as follows:

$$\begin{aligned}
\begin{pmatrix} \hat{\mathbf{H}}_i^t \\ \hat{\mathbf{H}}_i^m \end{pmatrix} &= \begin{pmatrix} \mathbf{Q}^t \\ \mathbf{Q}^m \end{pmatrix} \begin{pmatrix} \mathbf{K}^{t\top} & \mathbf{K}^{m\top} \end{pmatrix} \begin{pmatrix} \mathbf{V}^t \\ \mathbf{V}^m \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{Q}^t\mathbf{K}^{t\top}\mathbf{V}^t + \mathbf{Q}^t\mathbf{K}^{m\top}\mathbf{V}^m \\ \mathbf{Q}^m\mathbf{K}^{m\top}\mathbf{V}^m + \mathbf{Q}^m\mathbf{K}^{t\top}\mathbf{V}^t \end{pmatrix}.
\end{aligned} \tag{3}$$

Eq. 3 demonstrates that our cross-modal semantic concept enhancement module concurrently accounts for both intra-modal and inter-modal information. In practice, we will use multiple attention heads to enhance representation diversity.

2.4 Multi-scale Information Allocation Estimation

Fine-grained Allocation Estimation We first use class frequency to estimate class distribution $\pi_e^m = \frac{\mathbf{n}}{\sum_c n_c}$, where, $\mathbf{n}$ is the quantity of each category, n_c represents the number of category c in the training set. Based on this, we can obtain the debiased predictive distribution of a given sample x_i:

$$\mathbf{p}_i^m = \mathrm{softmax}(\mathrm{logits}(\hat{\mathbf{H}}_i^m) - \alpha \log \pi_e^m), \tag{4}$$

where, $\mathrm{logits}(\cdot)$ refers to the logits operation through prototype features $\hat{\mathbf{H}}_i^m$. We further define the positive score provided by the modality m conceptual prototype $\hat{\mathbf{H}}_i^m$ of sample x_i to the modality k conceptual prototype $\hat{\mathbf{H}}_j^k$ of sample x_j as $w_{ij}^{(m,k)} = \mathrm{Sim}(\mathbf{p}_i^m, \mathbf{p}_j^k)$, where, $\mathrm{Sim}(\cdot)$ denotes the cosine similarity function and $m, k \in \mathcal{M}$. Subsequently, we can obtain the distribution of allocation scores for $\mathbf{p}_i^m$ and $\mathbf{p}_j^k$ as :

$$p(w_{ij}^{(m,k)}) = \frac{w_{ij}^{(m,k)}}{\sum_b w_{ib}^{(m,k)}}. \tag{5}$$

Here, $p(w_{\hat{i}}^{(m,k)}|j)$ represents positive weight score between $\hat{\mathbf{H}}_i^m$ and $\hat{\mathbf{H}}_j^k$. For ease of reference, we define the fine-grained relationship weight matrix between modalities m and k as $W^{(m,k)} = \{w_{ij}^{(m,k)}|i,j \in [0,N]\}$. We utilize $W^{(m,k)}$ to balance the fine-grained information exchange between modalities m and k.

Coarse-grained Allocation Estimation The coarse-grained allocation estimation strategy directly affects the global modality information. We assign a learnable weight score $\pi_z^{(m,k)}$ to modalities m and k as the distribution estimate for global modality weighting:

$$\pi_z^{(m,k)} = \mathrm{softmax}\{z^{(m,k)}\}, k \in \mathcal{M}\backslash\{m\}, \tag{6}$$

where, $z^{(m,k)}$ represents a learnable parameters across modalities m and k. We utilize weight score $\pi_z^{(m,k)}$ to balance the coarse-grained information exchange between modalities m and k.

2.5 Self-Balancing Contrastive Learning

To effectively incorporate multimodal information into PLMs in a balanced manner, we expand traditional contrastive loss to a multimodal self-balancing contrastive loss based on the positive weight score $p(w_i^{(m,k)}|j)$:

$$\begin{aligned}
\mathcal{L}_i^{(m,k)} &= -\sum_{j=1}^{N} \mathbf{1}(y_i = y_j) p(w_{ij}^{(m,k)}) \log p_{ij}^{(m,k)}, \\
p_{ij}^{(m,k)} &= \frac{\exp(m_i \cdot k_j/\tau)}{\sum_{a=1}^{N} \exp(m_i \cdot k_a/\tau)},
\end{aligned} \tag{7}$$

where, m_i and k_j denotes the representations obtained by feeding $\hat{\mathbf{H}}_i^m$ and $\hat{\mathbf{H}}_j^k$ into a linear projector, τ is the temperature coefficient, $p_{ij}^{(m,k)}$ is the contrastive logit, and $m, k \in \mathcal{M}$. Further, we can obtain the contrastive loss between modalities m and k as:

$$\mathcal{L}^{(m,k)} = \sum_{i=1}^{N} \mathcal{L}_i^{(m,k)}. \tag{8}$$

Theorem 1. *Given m_i and k_j represent different modal conceptual prototypes of two samples x_i and x_j, the model θ is trained through the proposed self-balancing contrastive loss, then, the optimal contrastive logits $p_{ij}^{(m,k)^*}$ are approximately the distribution of allocation scores $p(w_{ij}^{(m,k)})$, i.e., the following equation approximately holds:*

$$p_{ij}^{(m,k)^*} = p(w_{ij}^{(m,k)}). \tag{9}$$

The proof of Theorem 1 can be found in Appendix A.2. Theorem 1 indicates that models trained using the proposed loss are capable of aligning the weights of multimodal information consistent with $p(w_{ij}^{(m,k)})$. In essence, this suggests that the proposed loss function, as mentioned in Eq. 7, can be effectively utilized to achieve self-balancing in the allocation of multimodal information between m and k.

2.6 Total Objective

Our overall loss comprises a weighted aggregation loss for prediction, coupled with a multi-scale self-balancing contrastive loss.

Weighted Aggregation Loss We can obtain the predicted distribution $p_m(y|x_i)$ of each category under different modality information guidance as:

$$p_m(y_i|x_i) = \text{softmax}(W^m \hat{\mathbf{H}}_i^m + b^m), \tag{10}$$

where, W^m is a fully connected matrix, b^m is a bias vector, and $m \in \mathcal{M}$ denotes the type of modality. Further, we adopt a learnable weight ρ^m as the late fusion over the final output distributions of multimodal models as:

$$p(y_i|x_i) = \sum_{m \in \mathcal{M}} \rho_m p_m(y|x_i), \tag{11}$$

where, ρ^m is a learnable normalized weight and $\sum_{m \in \mathcal{M}} \rho_m = 1$. Therefore, we can obtain the weighted aggregation loss for label prediction as:

$$\mathcal{L}_{wa} = \sum_{i=i}^{N} \text{CE}(y_i, p(y_i|x_i)), \tag{12}$$

where, $\text{CE}(\cdot)$ denotes the cross entropy loss.

Self-Balancing Contrastive Loss To achieve equilibrium in multimodal information processing during the training phase, we propose to formulate the total multi-scale self-balancing loss as follows:

$$\mathcal{L}_{msb} = \sum_{m \in \mathcal{M}} \sum_{k \in \mathcal{M} \setminus \{m\}} \frac{\pi_z^{(m,k)} \mathcal{L}^{(m,k)}}{\mid \mathcal{M} \mid (\mid \mathcal{M} \mid -1)}. \tag{13}$$

The multi-scale self-balancing loss, denoted as $\mathcal{L}_{msb}$, incorporates strategies for allocating multimodal information at both fine-grained and coarse-grained levels, thus achieving an adaptive balance in the modality information distribution during the training phase. Therefore, the overall loss is:

$$\mathcal{L} = \mathcal{L}_{wa} + \lambda \mathcal{L}_{msb}, \tag{14}$$

where λ denotes the trade-off hyperparameter.

3 Experiments

3.1 Experimental Setup

Datasets. We conduct experiments across three types of tasks:

- **Natural Language Understanding (NLU).** We evaluate our method over over SST-2 (Socher et al., 2013), QNLI (Rajpurkar et al., 2016), QQP(Chen et al., 2017) , MNLI (Williams et al., 2018), MRPC (Dolan and Brockett, 2005), and STS-B (Cer et al., 2017) datasets from GLUE benchmark (Wang et al., 2018).
- **Comprehension Question Answering.** We select SQuADv1.1 (Rajpurkar et al., 2016) and SQuADv2.0 (Rajpurkar et al., 2018) to evaluate our method on comprehension question answering task.
- **Text Generation.** In the realm of text generation, we utilized the CommonGen (Lin et al., 2020) as the dataset, which is a constrained text generation task focused on generative commonsense reasoning.

Baselines. We compare our approach with the pre-trained language models (PLMs), multimodal contrastive pre-trained models (MC-PTMs), and visually-augmented pre-trained language models (VA-PLMs). (1) PLMs: We utilize BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XL-Net (Yang et al., 2019), and T5 (Raffel et al., 2020)

Base Model	Method	Modality	SST-2	QNLI	QQP	MNLI	MRPC	STS-B	Avg.
BERT-base	+None	T	89.3	87.9	87.2	79.4	81.7	84.4	84.98
	+VOKEN	T+I	92.2	88.6	88.6	82.6	83.5	86.0	86.83
	+iACE	T+I	91.7	88.6	89.1	82.8	85.8	86.6	87.43
	+VAWI	T+I	92.4	89.1	89.7	83.0	85.6	86.9	87.78
	+Ours	T+I	93.4	90.1	90.9	83.6	86.8	87.4	88.70
	+Ours	T+I+A	93.8	91.1	91.2	84.6	86.9	88.0	89.27
	+Ours	T+I+A+V	**94.1**	**91.4**	**91.5**	**85.5**	**87.4**	**88.2**	**89.68**
RoBERTa-base	+None	T	89.2	87.5	86.2	79.0	81.4	85.4	84.78
	+VOKEN	T+I	90.5	89.2	87.8	81.0	87.0	86.9	87.06
	+iACE	T+I	91.6	89.1	87.9	82.6	87.7	86.9	87.06
	+VAWI	T+I	91.6	90.6	87.9	82.4	88.5	88.3	88.21
	+Ours	T+I	92.5	91.9	88.7	83.9	88.9	88.8	89.12
	+Ours	T+I+A	92.8	92.3	89.3	84.3	89.2	89.3	89.53
	+Ours	T+I+A+V	**93.2**	**92.5**	**89.7**	**85.0**	**89.6**	**89.7**	**89.95**

Table 1: A performance comparison of various methods on the GLUE benchmark, with the best results emphasized in **bold**. The term "+None" indicates the direct fine-tuning of the backbone without incorporating additional information. The abbreviations "T, I, A, V" denote four distinct modalities: text, image, audio, and video, respectively. We continuously add modal-specific concept prototype extractors to assess their influence on performance metrics. The results of VOKEN, iACE and VAWI on GLUE are reported from Lu et al. (2022) and Guo et al. (2022).

as the backbones, and directly fine-tune them as baselines. (2) MC-PTMs: We select CLIP (Radford et al., 2021), AudioCLIP (Guzhov et al., 2021) and CLIP-ViP (Xue et al., 2022). (3) VA-PLMs: VOKEN (Tan and Bansal, 2020b), iACE (Lu et al., 2022), and VAWI (Guo et al., 2022) are chosen as the main baselines in this paper. VOKEN introduces the visual information into PLMs by pretraining on retrieved images. iACE and VAWI infuse visual knowledge into PLMs through fine-tuning on retrieved or generated images and implicit image information, respectively.

Implementation Details. For all baselines, we follow the setting including hyperparameters with their papers and implement all methods based on Huggingface Transformers (Wolf et al., 2020). The hidden sizes for image, audio, and video concept prototypes are set to 512, 1024, and 512. We employ Adam as the optimizer with a weight decay of 0.01 and tune all models for 3 epochs. **More details and ablation studies can be found in Appendix A.3 and A.4, respectively**.

3.2 Main Experimental Results

Evaluation on NLU Tasks. We present our experimental results on six datasets in Table 1. Analysis of Table 1 yields several insights:

(1) Enriching PLMs with additional modality information significantly enhances their performance in NLP tasks. Both VA-PLMs (i.e., VOKEN, iACE, and VAWI) and our multimodal-augment PLMs outperform their corresponding PLM baselines. We attribute this to the incorporation of multimodal information that imparts general object knowledge (including color and shape) into PLMs.

(2) MaCSC of employing solely visual conceptual prototypes achieves superior performance compared to VA-PLM baselines. The performance of both VOKEN and iACE is surpassed by MaCSC of employing solely visual prototypes. This disparity may stem from the inclusion of extraneous noise and modality gaps. And MaCSC outperforms VAWI, with this advancement credited to our semantic enhancement and self-balancing strategies.

(3) As the variety of modal information incorporated into PLMs increases, there is a corresponding gradual improvement in the performance of PLMs. We can observe that the inclusion of more diverse sources of semantic conceptual information yields greater improvements in performance. In addition, our method outperforms all baselines in terms of PLM performance gains across all backbone networks It demonstrates the effectiveness and generality of MaCSC in injecting multimodal information into PLMs.

Evaluation on Comprehension Question Answering Tasks. As illustrated in Table 2, we evaluate the performance of our method on comprehen-

Base Model	Method	Modality	SQuAD v1.1 5% Acc.	SQuAD v1.1 5% F1	SQuAD v1.1 100% Acc.	SQuAD v1.1 100% F1	SQuAD v2 5% Acc.	SQuAD v2 5% F1	SQuAD v2 100% Acc.	SQuAD v2 100% F1
BERT-base	+None	T	61.9	72.7	80.2	87.7	54.8	57.9	72.1	75.2
	+Image	T+I	62.3	73.7	80.7	88.0	56.5	59.1	73.4	76.0
	+Ours	T+I	63.0	74.3	81.1	88.5	57.3	60.8	73.9	76.8
	+Ours	T+I+A	64.3	75.5	81.7	88.9	57.7	61.3	74.3	77.1
	+Ours	T+I+A+V	**65.5**	**76.2**	**83.4**	**90.2**	**59.1**	**62.4**	**75.5**	**78.4**
RoBERTa-base	+None	T	70.5	79.4	83.3	90.1	62.6	68.5	77.6	81.2
	+Image	T+I	70.8	79.9	83.4	90.3	63.3	69.1	78.0	81.7
	+Ours	T+I	71.5	80.7	83.9	91.0	64.2	69.7	78.7	82.5
	+Ours	T+I+A	71.8	81.2	84.3	91.5	65.0	70.6	79.5	83.3
	+Ours	T+I+A+V	**73.0**	**82.3**	**85.8**	**92.2**	**66.7**	**71.4**	**81.0**	**84.2**
XLNet-large	+None	T	73.4	83.5	84.8	91.4	68.8	72.0	79.4	82.6
	+Image	T+I	74.5	84.2	85.0	91.8	69.2	73.7	79.8	82.9
	+Ours	T+I	76.0	85.6	85.7	92.1	70.5	74.4	80.1	83.4
	+Ours	T+I+A	77.3	86.2	86.2	92.8	71.0	74.7	80.8	83.6
	+Ours	T+I+A+V	**79.6**	**87.3**	**86.9**	**93.5**	**72.6**	**75.8**	**82.0**	**84.8**

Table 2: A performance comparison of various methods on SQuAD v1.1 and SQuAD v2.0 datasets, with the best results emphasized in **bold**. The term "+None" indicates the direct fine-tuning of the backbone without incorporating additional information. "+Image" denotes we add retrieved images through Bing web search engines and utilize CLIP as image feature extractors. "T, I, A, V" denotes four distinct modalities: text, image, audio, and video.

sion question-answering tasks on SQuAD v1.1 and v2.0. We employ Bing Image Search[1] for image retrieval based on text input and utilize the CLIP image encoder to extract visual features as conceptual prototypes. Furthermore, we present results obtained using varying amounts of training data to assess the performance of different methods under a low-resource setting. Based on the results, we have the following detailed observations:

(1) The improvement in model performance by retrieving real images is not as significant as the gain by incorporating visual concept prototypes. This further demonstrates that inevitable noise images and modality gaps can adversely impact the representations and comprehension capabilities of PLMs. Our methodology eliminates the need for retrieval or generation of corresponding modality data, while simultaneously considering the self-balancing of multimodal information.

(2) The increase in the types of available modality prototypes will improve the performance of PLMs in question-answer tasks. In addition, our method consistently achieves performance gains on BERT, RoBERTa, and XLNet backbone networks, which further demonstrates the generality of our approach for different PLM architectures.

(3) In resource-constrained scenarios, our approach demonstrates a more substantial improvement. We randomly select 5% of the samples from the SQuAD training set to evaluate the performance under low-resource settings. We can observe that under low-resource settings, our method significantly boosts the performance compared to the results obtained with a full data set This enhancement is likely due to the capacity of our method to integrate balanced and beneficial modality information into PLMs, effectively mitigating the performance decline caused by the scarcity of data.

Method	BLUE-4	METOR	Rouge-L	CIDER
+None	36.2	32.7	59.3	17.7
+Image	35.8	32.1	59.0	17.6
+Ours w/o $\mathcal{L}_{mbs}$	**38.5**	**34.1**	**62.4**	**18.5**

Table 3: Performance comparison on CommonGen, with the best results emphasized in **bold**. We use T5-3b as the backbone. "+Image" denotes we add retrieved images through Bing web search engines and utilize CLIP as image feature extractors.

Evaluation on the Text Generation Task. To evaluate the universality of our approach on the text generation task, we implement it on the T5-3b and conduct experiments using the CommonGen

[1] https://learn.microsoft.com/en-us/azure/cognitive-services/bing-image-search/overview

Base Model	Method	SST-2	QNLI	QQP	MNLI	MRPC	STS-B	Avg.
CLIP	+None	72.4	72.6	70.2	69.1	73.9	75.1	72.22
AudioCLIP	+None	73.1	75.7	73.3	71.0	75.6	77.8	74.42
CLIP-ViP	+None	73.5	76.4	72.6	72.4	76.4	78.5	74.97
BERT-base	w/o Concept Enhancement	93.5	90.5	90.6	84.2	86.1	87.3	88.69
	w/o Self-balancing Loss	92.9	90.2	89.9	84.0	85.8	87.9	88.45
	+Full	**94.1**	**91.4**	**91.5**	**85.5**	**87.4**	**88.2**	**89.68**
RoBERTa-base	w/o Concept Enhancement	92.3	91.6	89.0	84.4	88.9	89.1	89.19
	w/o Self-balancing Loss	91.9	91.1	88.5	83.6	89.0	88.7	88.80
	+Full	**93.2**	**92.5**	**89.7**	**85.0**	**89.6**	**89.7**	**89.95**

Table 4: Ablation experiments on the GLUE benchmark. "+None" represents utilizing the text encoder of MC-PTMs combined with a classifier for prediction.

dataset, as shown in Table 3. Specifically, we insert enhanced representations of all conceptual prototypes after the output features of the text encoder in T5, while canceling self-balancing calibration. The results show that MaCSC consistently improves the performance of T5-3b across various metrics on the text generation task. Furthermore, we observe that employing images sourced from the Web search API can negatively impact text generation performance. This indicates that the text generation task is more sensitive to noise information.

3.3 Ablation Study

In this section, we conduct a series of ablation experiments to verify the effectiveness of our proposed components and strategies.

The Effect of the Semantic Concept Enhancement. We employ BERT-base and RoBERTa-base as the backbones and conduct experiments to study the impact of our semantic concept enhancement module as illustrated in Table 4. It can be seen that the performance of models significantly decreases without the semantic concept enhancement module to fusion modality information. This verifies the following: (1) the semantic concept enhancement module is capable of aggregating multimodal information sufficiently and effectively, and (2) the integration of intra-modal and inter-modal features plays a significant role in enhancing the expressive capabilities of PLMs.

The Effect of the Self-balancing Calibration. To evaluate the impact of our self-balancing calibration strategy on model performance, we remove it and presented the results in Table 4 Observations indicate that the absence of self-balancing loss significantly diminishes performance. This suggests that balancing information across modal knowledge is essential in multimodal-augmented PLMs. Additionally, this result corroborates the effectiveness of our self-balancing loss in harmonizing multimodal information and enhancing the expressive capabilities of PLMs.

Source of Concept Prototype	SQuAD v1.1		SQuAD v2	
	Acc.	F1	Acc.	F1
Gaussian Noise (0M)	83.4	90.3	77.3	80.8
BERT-base (110M)	83.7	90.8	78.1	81.9
RoBERTa-large (355M)	84.3	91.2	78.9	82.5
T5-3b (1500M)	84.5	91.3	79.6	83.6
CLIP* (63M)	**85.8**	**92.2**	**81.0**	**84.2**

Table 5: Performance comparison of different sources of concept prototypes in our approach on SQuAD datasets. We employ RoBERTa-base as the base model. "CLIP*" represents our complete approach, i.e., using the text encoders of CLIP, AudioCLIP and CLIP-ViP.

The Effect of the Source of Conceptual Prototypes. In this study, we employ three distinct PLMs (i.e., BERT-base, RoBERTa-large, and T5-3b, along with random Gaussian noise) as alternatives to our CLIP-based conceptual prototype extractor to assess the significance of multimodal representation. The results, as shown in Table 5, reveal that our approach achieves the most substantial performance gain, surpassing even that of T5-3b. This finding suggests that the performance improvement associated with the conceptual prototype is primarily due to the multimodal information it encapsulates, rather than merely the textual features offered by the text encoder. In essence, the performance gain brought by our method is indeed due to the multimodal information we extract. Furthermore, this also demonstrates the significant importance of additional modal knowledge in enhancing the expressiveness and inferential capabilities of

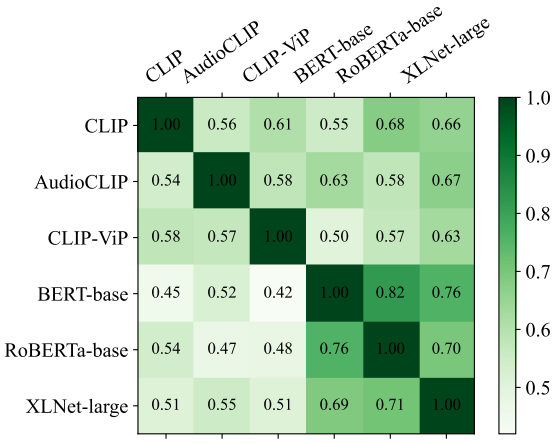


Figure 4: The overlap of correct predictions between each pair of models in the SQuAD v2 dataset.

PLMs.

Different Pre-trained Models Behave Differently. Following Yang et al. (2022), we mathematically define the concept of overlap in correct predictions between two models $\mathcal{M}_1$ and $\mathcal{M}_2$ as:

$$\mathcal{O}(\mathcal{M}_1, \mathcal{M}_2) = \frac{|\, \mathcal{S}_{\mathcal{M}_1} \bigcap \mathcal{S}_{\mathcal{M}_2} \,|}{|\, \mathcal{S}_{\mathcal{M}_1} \,|}, \qquad (15)$$

where, $\mathcal{S}_{\mathcal{M}}$ denotes the set of predictions made by model $\mathcal{M}$. We calculate the model overlap coefficients $\mathcal{O}$ among different models on the SQuAD v2 dataset in Figure 4. We can observe that MC-PTMs (i.e., CLIP, AudioCLIP, and CLIP-ViP) have a markedly smaller overlap with the other models. Conversely, the PLMs (i.e., BERT-base, RoBERTa-base, and XLNet-large) demonstrate a big mutual overlap. This difference explains the significant performance gain obtained through the integration of multimodal conceptual prototypes.

4 Related Works

Pre-trained Language Models. Recently, large-scale pre-trained language models (PLMs) have demonstrated remarkable success through self-supervised learning on extensive text corpus (Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020). By fine-tuning these PLMs, significant improvements are observed in downstream tasks including natural language processing (NLP), question answering, and text generation. Previous work has shown that language learners trained solely on textual data can exhibit biases and a lack of multimodal understanding of the objective world (Zhang et al., 2022; Yang et al., 2022; Cheng et al., 2023; Zhuang et al., 2024; Cheng et al., 2024). Furthermore, recent research suggests that these limitations are not mitigated by merely expanding the text corpus (Paik et al., 2021; Guo et al., 2022; Zhang et al., 2022; Zhu et al., 2024b,a). In this paper, we propose a universal multimodal-augmented framework for enriching the integration of multimodal semantics for PLMs.

Multimodal Contrastive Pre-trained Models. MC-PTMs are trained on a substantial corpus of modality-pairing samples, mapping multiple different modalities to a unified representation space. CLIP (Radford et al., 2021) performs pre-training on a large number of image text pairs, achieving alignment between image semantics and text semantics. AudioCLIP (Guzhov et al., 2021) extends CLIP to audio modality and CLIP-ViP (Xue et al., 2022) proposes a new strategy to transfer the image domain to the video domain. The latest work indicates that models similar to CLIP have the problem of modality gaps (Liang et al., 2022). Here, we utilize MC-PTMs to obtain text representations aligned with multimodal data as proxies,

Language Models with Additional Modal Enhancement. Most studies consider injecting visual modality information into PLMs through retrieval or image generation. Tan and Bansal (2020a); Wang et al. (2022) integrate visual knowledge during the pre-training of PLMs. Lu et al. (2022); Guo et al. (2023); Wang et al. (2024) fine-tune PLMs by introducing visual information during the fine-tuning phase. In addition, Yang et al. (2022) generates or retrieves images and encodes image representations to enhance zero-shot NLU. As a comparison, our approach is more efficient to leverage modality information including image, audio, and video in a balanced manner for enhancing PLMs.

5 Conclusion

In this paper, we propose a novel general multimodal-augmented framework with self-balancing calibration for PLMs called MaCSC. MaCSC can efficiently inject other modal semantic information into PLMs using multimodal conceptual prototypes. In addition, MaCSC adopts a novel self-balancing contrastive loss to achieve multi-scale self-balancing calibration of multimodal information during fine-tuning PLMs. Experimental results show that MaCSC consistently improves the performance of PLMs across various architectures and scales on multiple NLP tasks.

Ethics Statement and Limitation

Ethics Statement We conduct all experiments on the public datasets, which do not contain any offensive content or information with negative social impact. The focus of this article is to enhance the multimodal common sense of PLMs, and our model does not have uncontrollable outputs. Therefore, we ensure that our paper complies with ethical review guidelines.

Limitation This article proposes a new framework to balance and efficiently inject multimodal semantics into PLMs. However, the training of our proposed model relies on fine-tuning operations on PLMs, which may not be easily achievable in specific situations. Fully fine-tuning some super large language models may not be realistic. Therefore, exploring how to reduce fine-tuning costs and even provide training-free methods is more valuable for practical scenarios, which is also an important direction for future work. Furthermore, We use CLIP, AudioCLIP, and CLIP-ViP text encoders to generate modality-specific concepts as proxies. As pre-trained models, these MC-PTMs may also contain biases learned from the pre-trained corpus, which may lead to inappropriate bias predictions for some NLP tasks.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in technicolor. In *Annual Meeting of the Association for Computational Linguistics*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs. In *""*.

Xuxin Cheng, Bowen Cao, Qichen Ye, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023. ML-LMCL: Mutual learning and large-margin contrastive learning for improving ASR robustness in spoken language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6492–6505, Toronto, Canada. Association for Computational Linguistics.

Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yaowei Li, Xianwei Zhuang, and Yuexian Zou. 2024. Towards multi-intent spoken language understanding via hierarchical attention and optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17844–17852.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Conference on Automated Knowledge Base Construction*.

Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji rong Wen. 2022. Visually-augmented pretrained language models for nlp tasks without images. In *Annual Meeting of the Association for Computational Linguistics*.

Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji-Rong Wen. 2023. Visually-augmented pretrained language models for NLP tasks without images. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14912–14929, Toronto, Canada. Association for Computational Linguistics.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas R. Dengel. 2021. Audioclip: Extending clip to image, text and audio. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. 2022. Mind the

gap: Understanding the modality gap in multi-modal contrastive representation learning. *ArXiv*, abs/2203.02053.

Bill Yuchen Lin, Minghan Shen, Wangchunshu Zhou, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Findings*.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Imagination-augmented natural language understanding. *arXiv preprint arXiv:2204.08535*.

Tobias Norlund, Lovisa Hagström, and Richard Johansson. 2021. Transferring knowledge from vision to language: How to achieve it and how to measure it? In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Cory Paik, Stephane T Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2020a. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2020b. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2022. Visually-augmented language modeling. *arXiv preprint arXiv:2205.10178*.

Ziyang Wang, Heba Elfardy, Markus Dreyer, Kevin Small, and Mohit Bansal. 2024. Unified embeddings for multimodal retrieval via frozen LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1537–1547, St. Julian's, Malta. Association for Computational Linguistics.

Xiaoyan Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10938–10947.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Rui Song, Houqiang Li, and Jiebo Luo. 2022. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *International Conference on Learning Representations*.

Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2022. Z-LaVI: Zero-shot language solver fueled by visual imagination. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1186–1203, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022. Visual commonsense in pretrained unimodal and multimodal models. *ArXiv*, abs/2205.01850.

Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024a. Dance with labels: Dual-heterogeneous label graph interaction for multi-intent spoken language understanding. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1022–1031, New York, NY, USA. Association for Computing Machinery.

Zhihong Zhu, Xuxin Cheng, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2024b. Aligner²: Enhancing joint multiple intent detection and slot filling via adjustive and forced cross-task alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19777–19785.

Xianwei Zhuang, Xuxin Cheng, and Yuexian Zou. 2024. Towards explainable joint models via information theory for multiple intent detection and slot filling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19786–19794.

A Appendix

A.1 Theoretical Analysis of conceptual prototypes.

Theorem 2. *Assuming model θ is a CLIP-like MC-PTM trained on a substantial dataset comprising pairs of modalities k and t, let Z_k and Z_t denote representations generated through θ for k and t, respectively, and Y represents the label set for a specific NLP task, then maximizing the mutual information between Z_t and Y is equivalent to maximizing the mutual information between Z_k and Y, i.e,*

$$\max \mathbf{I}(Z_t, Y) \Leftrightarrow \max \mathbf{I}(Z_k, Y), \tag{16}$$

$\mathbf{I}(\cdot)$ represents mutual information.

Proof. Due to θ being an MC-PTM trained on a large amount of multimodal data pairs, the two highly aligned modal latent variables Z_t and Z_k are closely related. Therefore, we have the following inference:

$$p(Z_t|Z_k) \approx p(Z_t|Z_k) \approx 1. \tag{17}$$

Based on Eq. 17, we can derive $p(z_k, y) \approx p(z_t, y)$ from the total probability formula. According to the definition of mutual information, we have:

$$\begin{aligned} I\left(Z_t, Y\right) &= \sum_{z_t, y} p\left(z_t, y\right) \log \left(\frac{p\left(z_t, y\right)}{p\left(z_t\right) p(y)} \right); \\ I\left(Z_k, Y\right) &= \sum_{z_k, y} p\left(z_k, y\right) \log \left(\frac{p\left(z_k, y\right)}{p\left(z_k\right) p(y)} \right). \end{aligned} \tag{18}$$

In this case, if $\mathbf{I}(Z_t, Y)$ is maximized, $\mathbf{I}(Z_k, Y)$ will also be implicitly maximized due to the high correlation between Z_k and Z_t, and vice versa.

Claim 2 indicates that maximizing the information entropy between the intermediary proxy Z_t and Y is essentially in maximizing the information between Z_k and label Y. In other words, Z_t is a well-implicit conceptual prototype for modality k. We can utilize Z_t to efficiently transfer multimodal knowledge into PLMs to augment cognitive processing and understanding.

A.2 Proof of Theorem 1

Proof. We design a novel multimodal self-balancing contrastive loss between modalities m and k by incorporating a distribution of allocation

scores $p(w_{ij}^{(m,k)})$ as follows:

$$\mathcal{L}_i^{(m,k)} = -\sum_{j=1}^{N} \mathbf{1}(y_i = y_j) p(w_{ij}^{(m,k)}) \log p_{ij}^{(m,k)},$$
$$p_{ij}^{(m,k)} = \frac{\exp\left(m_i \cdot k_j / \tau\right)}{\sum_{a=1}^{N} \exp\left(m_i \cdot k_a / \tau\right)}, \tag{19}$$

where, m_i and k_j denotes the representations obtained by feeding $\hat{\mathbf{H}}_i^m$ and $\hat{\mathbf{H}}_j^k$ into a linear projector, τ is the temperature coefficient, $p_{ij}^{(m,k)}$ is the contrastive logit, and $m, k \in \mathcal{M}$. For simplicity, we define the distribution of allocation scores as $w_{ij} = p(w_{ij}^{(m,k)})$. Then, we approximate omit $\mathbf{1}(y_i = y_j)$ in Eq. 19 and solve the optimization problem with the Lagrange multiplier. The problem is defined as follows:

$$\begin{cases} \min & f(p_{i1}^{(m,k)}, \cdots, p_{in}^{(m,k)}) \\ & = -\sum_{j=1}^{n} \left(w_{ij} \cdot \log p_{ij}^{(m,k)} \right), \\ s.t. & g(p_{i1}^{(m,k)}, \cdots, p_{in}^{(m,k)}) \\ & = \sum_{j=1}^{n} p_{ij}^{(m,k)} - 1 = 0, \end{cases} \tag{20}$$

where, where n refers to the number of instances in the training batch. And the Lagrangian function of Eq. 20 can be formulated as:

$$\mathcal{L}(\lambda, p_{i1}^{(m,k)}, \cdots, p_{in}^{(m,k)}) = -\sum_{j=1}^{n} \left(w_{ij} \cdot \log p_{ij}^{(m,k)} \right) + \lambda \left(\sum_{j=1}^{n} p_{ij}^{(m,k)} - 1 \right) \tag{21}$$

and its corresponding partial derivatives are:

$$\frac{\partial \mathcal{L}}{\partial p_{ij}^{(m,k)}} = \lambda - \frac{w_{ij}}{p_{ij}^{(m,k)}} = 0; \tag{22}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{k=1}^{n} p_{ij}^{(m,k)} - 1 = 0. \tag{23}$$

From Eq. 22 and Eq. 23, the optimal value of contrastive logits is approximately as:

$$p_{ij}^{(m,k)*} = p(w_{ij}^{(m,k)}), \tag{24}$$

which concludes the proof for Theorem 1.

A.3 More training details.

In this work, we utilize CLIP (ViT-B/32)[2] (Radford et al., 2021), AudioCLIP (Full-Trained)[3] (Guzhov et al., 2021) and CLIP-ViP (base-patch32)[4] (Xue et al., 2022) text encoders to obtain conceptual prototypes of image, audio, and video modalities, respectively. The trade-off hyperparameter in Eq 14 is set to 0.2 and the number of attention heads in the semantic concept enhancement module is set to 8. We set the learning rate of 1e-4 on GLUE benchmark, 3e-5 on SQuAD v1.1 and SQuADv2.0 datasets, and 2e-5 on CommonGen dataset. In addition, we set the training batch size of 32 on GLUE benchmark and CommonGen dataset, and 12 on SQuAD v1.1 and SQuAD v2.0 datasets. We use grid search to determine the optimal hyperparameters mentioned above. We use Spearman's correlation as the metric on STS-B and the remaining GLUE tasks using accuracy as the metric. All experiments are conducted on 8 RTX 4090 GPUs.

A.4 More Ablation Experiments and Analysis.

Text Encoder in Prototype Extractor	SQuAD v1.1		SQuAD v2	
	Acc.	F1	Acc.	F1
BERT-base	83.7	90.8	78.1	81.9
BERT-base**	83.9	91.0	77.8	81.5
Non-fixed CLIP*	84.4	91.4	79.6	82.9
Fixed CLIP*	**85.8**	**92.2**	**81.0**	**84.2**

Table 6: Performance comparison of representations pre-trained using different pre-training data in our approach. We employ RoBERTa-base as the base model. "Non-Fixed CLIP*" represents using the text encoders of CLIP, AudioCLIP and CLIP-ViP as prototype extractors and set them as trainable modules. "BERT-base**" represents pre-training BERT-base on captions in CC3M.

The Effect of the Pre-trained Dataset. We notice some discrepancies between the textual data employed in MC-PTMs and the dataset used for PLMs pre-training. To evaluate the effect of pre-training textual data on model performance, we utilize captions from the CC3M dataset (Sharma et al., 2018) for masked pre-training of the BERT-base model. We utilize the retrained BERT-base as conceptual prototype extractors to conduct this ablation study. As shown in Table 6, the results

[2] https://github.com/openai/CLIP
[3] https://github.com/AndreyGuzhov/AudioCLIP
[4] https://github.com/microsoft/XPretrain/tree/main/CLIP-ViP

reveal that pre-training with captions does not enhance performance and even leads to a decline in performance on the SQuAD v2.0 dataset. This further indicates that the performance gains from our concept prototypes come not from the pre-trained textual captions but from multimodal data such as images, audio, and videos. In addition, we change the fixed prototype extractors into learnable modules and observe that learnable extractors result in performance degradation. This may be due to fine-tuning breaking the original semantic structure of MC-PTMs, thereby weakening the paired mapping assumption of Theorem 2.

Methods	MemoryColor	ColorTerm	ObjectShape
CLIP	27.3	24.9	19.8
BERT-base	25.1	26.7	31.5
RoBERTa-base	26.9	25.4	32.3
MaCSC	**35.2**	**34.3**	**34.5**

Table 7: Performance comparison on visual reasoning tasks, with the best results highlighted in **bold**. We use BERT-base as the backbone to evaluate our MaCSC.

Evaluation on Visual Reasoning Tasks. In this study, we employ datasets focused on reasoning about color and shape, specifically the MemoryColor (Norlund et al., 2021), ColorTerm (Bruni et al., 2012), and ObjectShape (Zhang et al., 2022) datasets, to evaluate the efficacy of our method in facilitating the transfer of visual knowledge. The outcomes of these evaluations are presented in Table 7. The findings derived from the analysis of these results indicate that our MaCSC significantly enhances the capacity of PLMs to comprehend object colors and shapes, thereby demonstrating the effectiveness of our approach in augmenting the multimodal comprehension capabilities of PLMs.