FSMR: A Feature Swapping Multi-modal Reasoning Approach with Joint Textual and Visual Clues

Anonymous ACL submission

Abstract

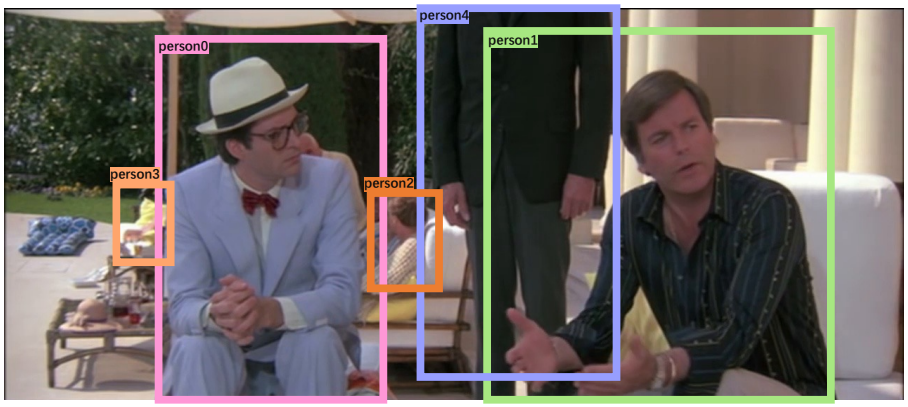
Multi-modal reasoning plays a vital role in bridging the gap between textual and visual information, enabling a deeper understanding of the context. This paper presents the Feature Swapping Multi-modal Reasoning (FSMR) model, designed to enhance multi-modal reasoning through feature swapping. FSMR leverages a pre-trained visual-language model as an encoder, accommodating both text and image inputs for effective feature representation from both modalities. It introduces a unique feature swapping module, enabling the exchange of features between identified objects in images and corresponding vocabulary words in text, thereby enhancing the model's comprehension of the interplay between images and text. To further bolster its multi-modal alignment capabilities, FSMR incorporates a multi-modal cross-attention mechanism, facilitating the joint modeling of textual and visual information. During training, we employ image-text matching and cross-entropy losses to ensure semantic consistency between visual and language elements. Extensive experiments on the PMR dataset demonstrate FSMR's superiority over state-of-the-art baseline models across various performance metrics.

1 Introduction

With the rise of social media, online news, and other multimedia platforms, textual information often coexists with information from other modalities like images. Multi-modal information processing has emerged as a crucial research direction in the field of natural language processing, regarded as a foundational and long-term task in both academia and industry (Yu et al., 2021; Chen et al., 2020).

Inspired by visual commonsense reasoning and textual inference, Dong et al. (2022) introduced the PMR (Premise-based Multi-modal Reasoning) dataset. In this task, models are required to use textual information (from the premise) and visual cues (from the image) to infer whether the hypothesis is true or not. Figure 1 illustrates an example from the PMR dataset. In this example, the model should recognize from the image that <person0> and <person1> are sitting together and conversing. Based on the textual premise, "<person0> and <person1> are talking about business", the model needs to determine whether the four hypotheses are true or not.

Compared to pure textual reasoning, multi-modal reasoning is more complex because it requires models to establish deep semantic connections between various modalities. To better fuse multi-modal inputs, researchers have designed self-supervised learning frameworks based on multi-modal encoders (Alberti et al., 2019; Li et al., 2019; Tan and Bansal, 2019). In recent years, given the outstanding performance of Pre-trained Language Models (PLM) in the field of natural language processing, many researchers have shown significant interest in Visual-Language Models (VLM) (Li et al., 2020). Although these methods have shown promise in reasoning tasks that heavily rely on visual cues, they still face challenges in aligning multi-modal data. For example, textual descriptions may zoom in on specific details of a scene, while the corresponding image may present an overall view of that scene. These differences can make it difficult for models to effectively merge information from two modalities when performing reasoning tasks. To address this problem, Li et al. (2023) designed a multi-modal contextual reasoning framework. Unlike traditional models, this framework incorporates a prefix for aligning images with text in pre-trained language models, enabling context semantic learning for both language and vision. However, the mentioned approaches do not delve into the fine-grained fusion of words in the premise and hypothesis with objects in the image, lacking granularity in multi-modal information integration.



Premise: <person0> and <person1> are talking about business.

Hypothesis:
A. <person0> and <person1> are sitting on the chair under the umbrella talking business.
B. <person0> and <person1> sit by the pool ready to go swimming as the sun goes down.
C. <person0> and <person1> sit in the pool ready to go swimming as the sun goes down.
D. <person0> and <person1> are sitting on the chair up above the umbrella talking business.

Figure 1: An example from PMR dataset

Our paper introduces a Feature Swapping Multi-modal Reasoning (FSMR) model for multi-modal reasoning.[1] The model utilizes a pre-trained visual-language model as an encoder, taking both text and image inputs to effectively represent features from both modalities. FSMR employs a unique feature swapping module that swaps the features of identified objects in the image with corresponding vocabulary words in the text, such as <person0> and <person1> in Figure 1. The swapped features are then incorporated into a prompt template and input into the language model, allowing the model to understand the context information fused between images and text. To further enhance the multi-modal alignment capability, FSMR introduces a multi-modal cross-attention mechanism, enabling joint modeling of textual and visual information. We adopt image-text matching loss and cross-entropy loss during training to ensure semantic consistency between vision and language. Extensive experiments on the standard PMR dataset for multi-modal reasoning validate the approach. Results demonstrate that FSMR outperforms state-of-the-art baseline models across various performance metrics.

The main contributions of this paper can be summarized as follows: (1) We introduce the Feature Swapping Multi-modal Reasoning (FSMR) model to address multi-modal reasoning tasks; (2) We introduce a multi-modal cross-attention mechanism that allows for joint modeling of textual and visual information; (3) To substantiate our approach, we conducted extensive experiments on the PMR dataset. These experiments clearly demonstrate that FSMR surpasses state-of-the-art baseline models across various performance metrics.

2 Related Work

To better fuse multi-modal inputs, researchers have designed self-supervised learning frameworks based on multi-modal encoders. Specifically, depending on the construction of the encoder, these multi-modal learning frameworks can be categorized into two types. The first framework uses a unified encoder to directly process multi-modal inputs (Sun et al., 2019; Alberti et al., 2019; Li et al., 2019). The second one initially employs two separate encoders to process textual and image data independently and then uses a joint encoder to integrate the representations obtained from both, achieving the goal of merging multi-modal information (Lu et al., 2019; Tan and Bansal, 2019). Among these, Cui et al. (2020) introduced a multi-modal alignment contrastive learning decoupled network. This approach introduces multi-modal contrastive losses between the text encoder and the image encoder, ensuring a high semantic match between the textual description and the corresponding image.

In recent years, given the outstanding perfor-

[1]The code is available at https://anonymous.4open.science/r/FSMR-8CED.

mance of Pre-trained Language Models (PLMs) in the field of natural language processing, many researchers have shown significant interest in Visual-Language Models (VLMs) (Krojer et al., 2022; Li et al., 2020; Wang et al., 2022). Lu et al. (2019) introduced a pre-trained model called VL-BERT for visual-language tasks. This model extends the Transformer encoder to accept visual and textual features as inputs. Still, in this process, context learning based on the multi-modal semantics of language and vision is often overlooked. To address this problem, Li et al. (2023) designed a multi-modal contextual reasoning framework. However, the mentioned approaches do not delve into the fine-grained fusion of words in the premise and hypothesis with objects in the image, lacking granularity in multi-modal information integration. In this paper, we proposes a Feature Swapping Multi-modal Reasoning (FSMR) model for multi-modal reasoning to tackle this problem.

3 Architecture

The overall structure of the FSMR model is depicted in Figure 2. The FSMR model utilizes a pre-trained Visual Language Model as an encoder to obtain representations of text and images. We introduce a feature swapping layer that swaps the features of objects in the image with corresponding word representations in the text. After obtaining these new representations, they are filled into a pre-designed prompt template and fed into a language model to compute cross-entropy loss. Additionally, FSMR incorporates a multi-modal multi-head attention module to integrate information from both text and images. The model employs an image-text matching loss to align text and image representations in the semantic space.

3.1 Encoder

In the PMR taspremise-based multi-modal reasoning task, each instance consists of two sentences (premise and hypothesis), an image denoted as V, and a label representing the relationship between the sentences (entailment or contradiction). The image input is denoted as V. An instance in a batch, denoted as $\mathcal{I}$, is represented as $(X^{(p)}, X^{(h)}, V, y)_i$, where $i = \{1, \ldots, K\}$ is the sample index, and K is the batch size. The goal is to learn a mapping function f on the training data, which predicts the category y based on the input.

The FSMR model utilizes the pre-trained ViLBERT as its encoder. To process complex inputs that combine text and images, FSMR employs a special concatenation method. The format is as follows: "`[CLS]` $X^{(p)}$ `[SEP]` $X^{(h)}$ `[IMG]` V". The embeddings for text and images are obtained from the encoder:

$$\boldsymbol{h}_{\text{CLS}}, \boldsymbol{w}_i, \boldsymbol{v}_j, \boldsymbol{h}_{\text{IMG}} = \text{ViLBERT}(X^{(p)}, X^{(h)}, V) \quad (1)$$

Where $\{\boldsymbol{w}_i | i = 1, 2, \cdots, n\}$ represents word representations in the premise and hypothesis. , with $n = l_1 + l_2$. $\{\boldsymbol{v}_j | j = 1, 2, \cdots, m\}$ represents objects in the image, with m denoting the number of objects. For `[CLS]` and `[IMG]`, the encoder outputs are denoted as $\boldsymbol{h}_{\text{CLS}}$ and $\boldsymbol{h}_{\text{IMG}}$, respectively, representing the overall representations for text and image.

3.2 Feature Swapping Layer

In order to enhance the model's understanding of multi-modal contexts, we introduce an innovative mechanism called Feature Swapping. For each object in the image, when a corresponding word is mentioned in the text, the embeddings representing that object in the image and the corresponding word in the text are swapped. Figure 3 displays an image with objects outlined in boxes. This image contains 11 objects (person0, person1, person2, person3, person4, tie, chair, chair, chair, chair, chair). Both "person0" and "person1" are mentioned in both the premise and hypothesis.

For the example in Figure 2, the hypothesis describes that "person0" and "person1" are discussing business, which aligns with the meaning described in the premise. The goal of the Feature Swapping Layer is to ensure that the model correctly aligns words in the text with corresponding image objects in the semantic space. The exchanged embeddings for word representations $\{\boldsymbol{w}_i | i = 1, 2, \cdots, n\}$ and object representations $\{\boldsymbol{v}_j | j = 1, 2, \cdots, m\}$ are denoted as $\boldsymbol{h}_w$ and $\boldsymbol{h}_v$, respectively:

$$\boldsymbol{h}_w = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{i-1}, [\boldsymbol{v}]_j, \boldsymbol{w}_{i+1}, \ldots, \boldsymbol{w}_n) \quad (2)$$

$$\boldsymbol{h}_v = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{j-1}, [\boldsymbol{w}]_i, \boldsymbol{v}_{j+1}, \cdots, \boldsymbol{v}_m) \quad (3)$$

In the above equations, $[\boldsymbol{v}]_j$ and $[\boldsymbol{w}]_i$ represent the swapped features $\boldsymbol{v}_j$ and $\boldsymbol{w}_i$, respectively, and their corresponding words and objects actually represent the same entity.

For the overall text representation $\boldsymbol{h}_{\text{CLS}}$ and the overall image representation $\boldsymbol{h}_{\text{IMG}}$, FSMR designs an aligner module to tightly integrate them, forming a fused representation of the image and text.

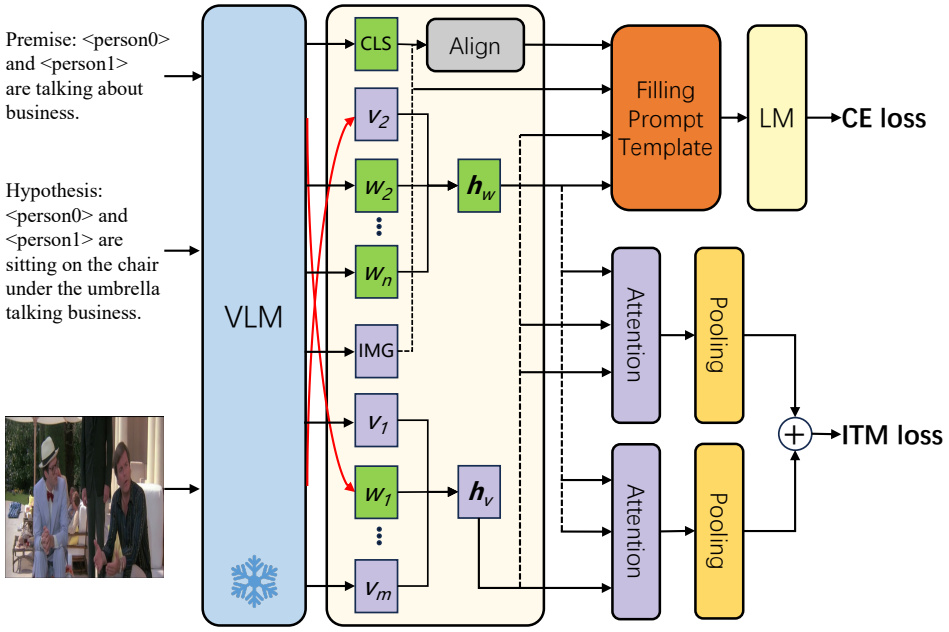


Figure 2: Overall Architecture of the FSMR Model

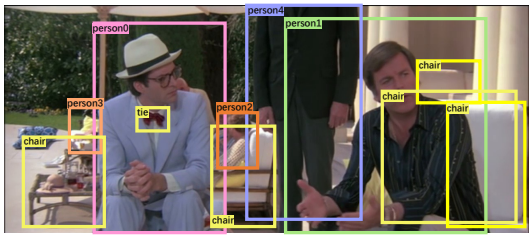


Figure 3: Example of objects in the image

This aligner module is not complex, consisting of linear layers and the Tanh activation function. The fused representation $\boldsymbol{A}$ is calculated by:

$$\boldsymbol{A} = \tanh(\boldsymbol{W} * \text{concat}(\boldsymbol{h}_{\text{CLS}}, \boldsymbol{h}_{\text{IMG}}) + \boldsymbol{b}) \quad (4)$$

$\boldsymbol{W}$ and $\boldsymbol{P}$ are trainable parameters used for linear transformation.

3.3 Prompt Template

After obtaining the aforementioned embedding representations, this section employs the widely adopted technique of prompt engineering to integrate the encoded information and fill it into a carefully designed prompt template. The predefined prompt template is as follows: "`[CLS]` Given an image with feature <$\boldsymbol{h}_{\text{IMG}}$>, the alignment feature is <$\boldsymbol{A}$>, objects identified as <$\boldsymbol{h}_v$> `[SEP]` <$\boldsymbol{h}_w$>".

This template is then input into a pre-trained language model (RoBERTa (Liu et al., 2019)). By constructing such prompt templates, existing image representations are embedded into the language model, transforming the multi-modal reasoning task paradigm into a purely language model reasoning paradigm. The output of RoBERTa's `[CLS]` representation, denoted as $\boldsymbol{S}_{\text{CLS}}$, is used for inference.

3.4 Multi-Head Attention Module

To effectively fuse language and visual information, this model introduces a multi-head attention module after the feature fusion layer. Given intermediate representations for vision and language, denoted as $\boldsymbol{h}_v$ and $\boldsymbol{h}_w$ respectively, separate linear layers are used to compute the query, key, and value matrices. In the traditional way, the query, key, and value matrices all originate from the same input. However, we adopt a cross-modal multi-head attention mechanism in FSMR. Specifically, the key and value matrices for the language modality are provided to the multi-head attention component for the vision modality as input, and vice versa, the key and value matrices for the vision modality are provided to the multi-head attention component for the language modality. The representations output by the two multi-head attention components are denoted as $\boldsymbol{O}_w$ and $\boldsymbol{O}_v$, computed as follows:

$$\boldsymbol{O}_w = \text{Multi-Head}(Q_w, K_v, V_v) \quad (5)$$

$$\boldsymbol{O}_v = \text{Multi-Head}(Q_v, K_w, V_w) \quad (6)$$

After obtaining the outputs of the two multi-head attentions, dimension reduction and capture of their main features are achieved first through a pooling layer (e.g., average pooling or max pooling). Let

$\boldsymbol{P}_w$ and $\boldsymbol{P}_v$ be the representations of $\boldsymbol{O}_w$ and $\boldsymbol{O}_v$, respectively, after pooling processing:

$$\boldsymbol{P}_w = \text{Pooling}(\boldsymbol{O}_w) \quad (7)$$

$$\boldsymbol{P}_v = \text{Pooling}(\boldsymbol{O}_v) \quad (8)$$

Next, the two pooled representations are concatenated to obtain the overall multi-head attention representation $\boldsymbol{S}_{\text{attn}}$:

$$\boldsymbol{S}_{\text{attn}} = \text{Concat}(\boldsymbol{P}_w, \boldsymbol{P}_v) \quad (9)$$

Pooling represents a pooling function (e.g., average pooling, max pooling, etc.), and Concat denotes the vector concatenation operation.

3.5 Objective Function

Image-Text Matching Loss In this section, we introduce the image-text matching loss function specifically designed for FSMR, aiming to ensure the effective alignment of visual and textual information. This loss function is denoted as $\mathcal{L}_{\text{ITM}}$. After obtaining the overall representation $\mathbf{S}_{\text{attn}}$ from multi-head attention, it is first passed through a linear layer. Subsequently, it is transformed into a probability p_{ITM} within the range of [0,1] using the sigmoid activation function:

$$p_{\text{ITM}} = \text{sigmoid}(\mathbf{W} \cdot \mathbf{S}_{\text{attn}} + \mathbf{b}) \quad (10)$$

$\mathbf{W}$ and $\mathbf{b}$ are trainable parameters, and y is the ground truth label for the example. Next, we calculate the loss function $\mathcal{L}_{\text{ITM}}$ as follows:

$$\mathcal{L}_{\text{ITM}} = -\left(y \log p_{\text{ITM}} + (1-y)\log(1-p_{\text{ITM}})\right) \quad (11)$$

Cross-Entropy Loss In addition to the image-text matching loss, the `[CLS]` representation $\boldsymbol{S}_{\text{CLS}}$ generated by RoBERTa utilizes a softmax-based Cross-Entropy loss function for classification:

$$\mathcal{L}_{\text{CE}} = CrossEntropy(\boldsymbol{W} \cdot \boldsymbol{S}_{\text{CLS}} + \boldsymbol{b}, y) \quad (12)$$

$\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters, and y represents the annotated label for this example.

Overall Loss Function The overall training objective of the FSMR model, denoted as $\mathcal{L}$, is the weighted average of the cross-entropy loss and the image-text matching loss, represented as:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{ITM}} \quad (13)$$

α and β are hyperparameters used to balance the loss functions.

4 Experimental Setup

4.1 Benchmark Dataset

To validate the effectiveness of the proposed model, experiments were conducted on the high-quality PMR dataset (Dong et al., 2022). These samples were created through a multi-stage crowd-sourcing process. Crowd-workers, guided by predefined categories, selected high-quality movie screenshots and manually curated premise templates to write a genuine hypothesis along with three distractor options in a cross-checking procedure, based on the provided premise and the image. Classification accuracy is used as the evaluation metric in the experiments.

4.2 Implementation Details

The model is implemented using PyTorch. We utilize Faster R-CNN (He et al., 2017) as the image feature encoder for extracting visual regions. For visual-linguistic alignment, we employ Oscar as the visual language aligner, and RoBERTa serves as the multi-modal context network. The training details can be found in Appendix A.

4.3 Baseline Models

We compare FSMR with pre-trained language models and multi-modal models as follows:(1)BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are large-scale language models based on the Transformer architecture; (2)ViLBERT (Lu et al., 2019) is a cross-modal pre-trained model with dual data streams; (3)ERNIE-VL (Yu et al., 2021) uses a single-stream fusion encoder and leverages structured knowledge obtained to learn joint representations; (4)UNITER (Chen et al., 2020) integrates visual information and utilizes joint multi-modal embeddings to support heterogeneous downstream visual language tasks; (5) Oscar (Li et al., 2020) is a single-stream fusion encoder model that simplifies alignment learning by using object labels detected in images as anchors; (6) OFA (Wang et al., 2022) is a sequence-to-sequence cross-modal learning framework that unifies various cross-modal and uni-modal tasks; (7) MVPTR (Li et al., 2022) is a pre-trained cross-modal model that introduces multi-level semantic alignment between vision and language; (8) CALeC (Yang et al., 2022) is a unified prediction and generation model for certain visual-language tasks; (9) PromptFuse (Liang et al., 2022) is a prompt-based learning approach to incorporate visual information into language models;

(10) ModCR (Li et al., 2023) is a multi-modal contextual reasoning framework that incorporates a prefix capable of learning alignment between images and text into pre-trained language models.

5 Experiment Results

5.1 Main Results

Method ↓ Types →	Validation	Testing
BERT-B	-	65.2
VL-BERT-B	-	75.4
ERNIE-VL-B	-	79.0
UNITER-B	-	77.4
Oscar-B	77.7	76.1
RoBERTa-L	77.3	75.0
PromptFuse	77.4	76.5
VL-BERT-L	-	79.3
ERNIE-VL-L	-	79.9
UNITER-L	-	77.0
OFA-L	79.9	79.1
MVPTR	79.5	78.9
CALeC	80.1	78.7
ModCR	<u>85.0</u>	<u>83.6</u>
FSMR	**86.4**	**84.8**

Table 1: Model Performance (accuracy) on the PMR dataset. The results of BERT, VL-BERT, ERNIE-VL and UNITER are reported by Dong et al. (2022). For baselines, "B" and "-L" indicate the base and large version, respectively. The underscore and bold indicate the second highest value and best performance(same as following tables).

We conducted experiments on the PMR dataset to evaluate the model's performance. Table 1 displays the results of FSMR and other baseline models on both the validation and test sets. All results are the averages of five runs with different random seeds, and the best results are highlighted in bold. Some models, such as BERT-B, VL-BERT-B, ERNIE-VL-B, VL-BERT-L, and UNITER-L, were evaluated only on the test set, and validation set data were not provided in the original work.

From the test set data, it is evident that most models perform in the range of 75% to 80%. This demonstrates that the multi-modal natural language reasoning task on the PMR dataset is indeed challenging. FSMR excels, achieving the best performance on both the validation and test sets, with accuracy rates of 86.4% and 84.8%, respectively, significantly outperforming other baseline models. Compared to the state-of-the-art baseline model ModCR, FSMR exhibits a substantial improvement, increasing accuracy by 1.4% on the PMR validation set and 1.2% on the test set. This improvement is relatively significant in natural language processing tasks. The performance of BERT-B and RoBERTa (text input only) suggests that reasoning based solely on the premise text can lead to correct choices, but with lower accuracy. FSMR, using RoBERTa-L as its primary backbone, outperforms pre-trained VLM and LM models on both datasets. This indicates that the FSMR approach effectively integrates semantic information from different modalities when performing inference.

Method ↓ Types →	AT↑	D1↓	AF↓	D2↓
BERT-B	65.2	19.8	19.6	4.5
Oscar-B	76.1	10.2	12.1	1.7
RoBERTa-L	75.0	17.7	6.1	<u>1.2</u>
PromptFuse	76.5	16.5	<u>5.9</u>	<u>1.2</u>
ERNIE-VL-L	79.9	10.7	8.2	<u>1.2</u>
OFA-L	79.1	9.7	9.9	1.3
MVPTR	78.9	**7.5**	11.8	1.8
CALeC	78.7	8.6	10.9	1.8
ModCR	<u>83.6</u>	9.2	**5.6**	1.6
FSMR	**84.8**	<u>8.4</u>	<u>5.9</u>	**0.9**

Table 2: Detailed performance on the test set of PMR. The results of BERT and ERNIE-VL are reported by (Dong et al., 2022). AT, D1, AF, D2 represent the Action True and Image True, Action True yet Image False, Action False yet Image True, Action False and Image False, respectively. "Action True or False" indicate the answer whether meets the premise. Similarly, "Image True or False" show the answer whether meets the image information.

Table 2 provides a comprehensive overview of the model's performance on the PMR test set, aiming to evaluate the model's accuracy in reasoning across different types of answer candidates. The table presents the model's reasoning distribution across these categories, allowing for an in-depth analysis of potential factors contributing to classification errors—whether they are due to semantic disparities or deviations in image information. Observing the table, FSMR exhibits superior overall performance compared to other baseline models, with error rates of 8.4%, 5.9%, and 0.9% in D1, AF, and D2, respectively. Particularly in the D2 category, FSMR outperforms all other models. By combining RoBERTa as the context encoder for prompts, FSMR successfully achieves precise

alignment of textual and image semantics through feature swapping and cross-modal multi-head attention mechanisms. This not only retains robust text reasoning abilities, as indicated by the AF results, but also significantly enhances the utilization of image information, as shown in the D1 results. In summary, adding a visual-language semantic alignment mechanism to vision-augmented language models is crucial. Moreover, there is still room for optimization in the area of contextual reasoning in current Vision-Language Models.

5.2 Ablation Study

Model	Validation	Testing
FSMR	**86.3**	**84.8**
-Feature Swapping	85.8	84.4
-Prompt Template	84.5	83.4
-Multi-head Attention	85.3	83.9
-ITM loss	85.5	84.1
-CE loss	85.4	84.2

Table 3: The ablation results of FSMR on the test set of PMR.

To gain a better understanding of the contributions of each key component within FSMR, we conducts ablation studies on the PMR dataset. The results are presented in Table 3. Notably, removing the feature swapping layer results in a performance drop of 0.5% on the validation set and 0.4% on the test set, emphasizing its importance in enhancing alignment between image objects and textual words. Removing the prompt template has a more significant impact, causing accuracy to decrease by 1.8% on the validation set and 1.4% on the test set. This demonstrates that the prompt template plays a crucial role in incorporating image information into the language context, which is vital for reasoning accuracy. The removal of the multi-head attention module leads to a substantial performance drop, with accuracy decreasing by 1.0% on the validation set and 0.9% on the test set. This highlights the critical role of the multi-head attention module in aligning and fusing textual and visual information effectively. Removing the image-text matching loss alone results in a decrease of 0.8% on the validation set and 0.7% on the test set, underscoring its positive impact on training the model to align image and text information. Finally, the removal of the cross-entropy loss has a relatively smaller impact, causing a decrease of 0.9% on the validation set and 0.6% on the test set.

5.3 Analysis of Feature Swapping

Method	Val	Test
Unidirectional (Image to Text)	85.9	84.4
Unidirectional (Text to Image)	84.7	83.8
Bidirectional	**86.3**	**84.8**
Hybrid	85.4	83.7

Table 4: Experimental Results with Different Feature Swapping Strategies

Table 4 illustrates the experimental results of different feature swapping methods within the Feature Swapping Layer. When replacing text features with image features, the model achieves validation and test set accuracy of 85.9% and 84.4%, respectively, which exhibit a relatively modest decrease compared to bidirectional swapping. However, when replacing image features with text features, the model's performance is notably lower, with validation and test set accuracy of 84.7% and 83.8%, respectively.

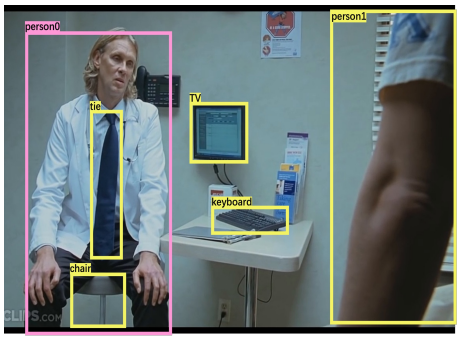
The model performs exceptionally well with bidirectional feature swapping, achieving validation and test set accuracy of 86.3% and 84.8%, respectively. In the case of hybrid swapping, which involves randomly choosing one of the four methods (unidirectional image, unidirectional text, bidirectional, or no swapping), the model's performance is slightly lower than bidirectional swapping but falls between the two unidirectional methods. The accuracy on the validation and test sets for hybrid swapping is 85.4% and 83.7%, respectively, indicating that the hybrid swapping strategy indeed leverages some of the advantages of bidirectional swapping but may not consistently achieve optimal performance under all conditions.

5.4 Analysis of Multi-Head Attention

Strategy	Validation	Testing
Visual Attention	86.1	84.1
Language Attention	84.5	82.8
Mixed Attention	**86.3**	**84.8**

Table 5: Experimental Results with Different Multi-Head Attention Strategies

As shown in Table 5, this section analyzes the impact of different multi-head attention strategies



Premise: [person0] is thought to be a responsible doctor.

Prediction: ModCR: D1, FSMR: AT.

Hypothesis:

AT. Dressed in a white coat, [person0] is in the exam room telling his patient [1] what to do and reminding him that he must come back regularly.

AF. It is nearly time to go home, so [person0], in his white coat, is perusing his last patient in the exam room. He doesn't ask him about his illness but just gives him some medicine casually.

D1. Dressed in a blue coat, [person0] is in the exam room telling his patient [1] what to do and reminding him that he must come back regularly.

D2. It is nearly time to go home , so [person0], in his blue coat, is perusing his last patient in the exam room. He doesn't ask him about his illness but just gives him some medicine casually.

Figure 4: A case from the PMR Test Set. Blue indicates content that contradicts the textual premise, while yellow marks content inconsistent with the image. Green and red emoticons signify correct and incorrect options, respectively.

on model performance.

From the table, it can be observed that when the model uses visual modality attention only, it achieves an accuracy of 86.1% on the validation set and 84.1% on the test set. In contrast, when using language modality attention only, the model's accuracy on the validation and test sets is 84.5% and 82.8%, significantly lower than the pure visual modality strategy. When both visual and language modality attention mechanisms are used simultaneously, the model's accuracy on the validation and test sets surpasses that of single modality strategies, reaching 86.3% and 84.8%, respectively. This demonstrates that combining visual and language information leads to better performance and underscores the importance of multi-modal attention in understanding and integrating modality information.

5.5 Case Analysis

A case from the PMR test set is illustrated in Figure 4. In this case, the textual premise states that "[person0] is considered a responsible doctor," and in the image, [person0] is seen wearing a white coat while sitting in a chair. Among the four options, the 'AT' option conveys that "[person0] is wearing a white coat, providing guidance in the examination room and reminding the patient to return for a follow-up," which aligns with both the image and the textual premise. The 'AF' option suggests that "[person0] is about to finish work, did not inquire about the patient's condition, and casually provided some medication," which contradicts the responsible doctor mentioned in the premise. However, in the 'D1' option, the blue coat contradicts the image information. The 'D2' option combines elements from both 'AF' and 'D1' and is inconsistent with both the textual and image information.

For this example, the baseline ModCR model's inference results in 'D1,' indicating that this model failed to effectively integrate image information for reasoning and did not recognize the contradiction between the answer and the image content. In contrast, FSMR can jointly model multi-modal information to infer the correct answer, identifying inconsistencies with both the image and textual premise. This demonstrates that FSMR, through multi-modal attention mechanisms and alignment loss, can fuse and comprehend textual and image data, enabling cross-modal contextual semantic reasoning.

6 Conclusion

We propose a Feature Swapping Multi-modal Reasoning model named FSMR. The features that are swapped are subsequently integrated into a prompt template and fed into a language model.

To further enhance the alignment and complementarity between text and images, FSMR introduces a multi-modal cross-attention mechanism, which plays a pivotal role in deepening the integration of visual and language information. Additionally, the model's training strategy is meticulously designed, ensuring that FSMR effectively aligns and integrates visual and textual information in the context of multi-modal reasoning tasks. Experimental evaluations demonstrate FSMR's superior performance on the standard PMR dataset. Furthermore, we delves into a comprehensive exploration and analysis of the components of the FSMR model.

7 Limitations

The FSMR model exhibits promising advancements in multi-modal reasoning, but certain limitations should be considered. Its performance heavily relies on diverse and high-quality training data, and generalization to different domains beyond the PMR dataset may be a challenge. Additionally, while superior on the PMR dataset, FSMR's performance on other multi-modal datasets remains unexplored. Addressing these issues is crucial for enhancing the model's practical applicability across various multi-modal reasoning tasks.

References

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2131–2140.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Wanyun Cui, Guangyu Zheng, and Wei Wang. 2020. Unsupervised natural language inference via decoupled multimodal contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5511–5520, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Qingxiu Dong, Ziwei Qin, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, et al. 2022. Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 932–946.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020*, pages 121–137. Springer.

Yunxin Li, Baotian Hu, Chen Xinyu, Yuxin Ding, Lin Ma, and Min Zhang. 2023. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10757–10770, Toronto, Canada. Association for Computational Linguistics.

Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. 2022. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405.

Sheng Liang, Mengjie Zhao, and Hinrich Schütze. 2022. Modular and parameter-efficient multimodal fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2976–2985.

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 preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, pages 13–23.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.

Tijmen Tieleman and Geoffrey Hinton. 2012. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *COURSERA Neural Networks Mach. Learn*, 17.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Qian Yang, Yunxin Li, Baotian Hu, Lin Ma, Yuxin Ding, and Min Zhang. 2022. Chunk-aware alignment and lexical constraint for visual entailment with natural language explanations. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3587–3597.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

A Training Details

During the model training process, we employ the RMSprop optimizer (Tieleman and Hinton, 2012). We train the model for 30 epochs with a batch size of 8. The base learning rate of the model is set to 4e-06, with a weight decay of 8e-05, ϵ set to 5e-05, and it is adjusted using a linear scheduler. To ensure that the processed sequence information does not exceed the model's capacity, we set the maximum sequence length to 150. The length of the visual prefix is set to 3, while the cross-modal alignment prefix is set to 5. The number of heads in the multi-modal multi-head attention module in the model is set to 16, with a dropout rate of 0.2. All experiments are conducted 5 times using different random seeds, and the average results are reported. All methods select the best-performing model using the validation set. To ensure efficient computation, all experiments are carried out on GeForce GTX 3090Ti.