

# Enhancing Semantics in Multimodal Chain of Thought via Soft Negative Sampling

Guangmin Zheng, Jin Wang\*, Xiaobing Zhou, Xuejie Zhang

School of Information Science and Engineering

Yunnan University

Kunming, China

gmzheng@mail.ynu.edu.cn, {wangjin, zhouxb, xjzhang}@ynu.edu.cn

## Abstract

Chain of thought (CoT) has proven useful for problems requiring complex reasoning. Many of these problems are both textual and multimodal. Given the inputs in different modalities, a model generates a rationale and then uses it to answer a question. Because of the hallucination issue, the generated soft negative rationales with high textual quality but illogical semantics do not always help improve answer accuracy. This study proposes a rationale generation method using soft negative sampling (SNSE-CoT) to mitigate hallucinations in multimodal CoT. Five methods were applied to generate soft negative samples that shared highly similar text but had different semantics from the original. Bidirectional margin loss (BML) was applied to introduce them into the traditional contrastive learning framework that involves only positive and negative samples. Extensive experiments on the ScienceQA dataset demonstrated the effectiveness of the proposed method. Code and data are released at <https://github.com/zgMin/SNSE-CoT>.

**Keywords:** Multimodal chain of thought, Soft negative sampling, Bidirectional margin loss.

## 1. Introduction

Artificial intelligent systems have long been aimed at behaving dependably and learning complicated tasks quickly. As humans, we can use an explicit chain-of-thought (CoT) reasoning process, which is often articulated as an explanation, to make dependable decisions (Wei et al., 2022). Nevertheless, for a specific task, machine learning models are often trained using a large number of input-output samples. These black-box approaches only produce a final decision without consistently disclosing the underlying reasoning. CoT methods have recently been demonstrated to be extremely useful for large-scale language models (LLMs) in handling tasks that require complex reasoning. Most previous studies have focused only on language modalities, whereas inference may exist in multiple modalities, such as visual question answering (VQA). Given the inputs in different modalities, an intelligent system is required to infer answers using multi-hop intermediate reasoning.

Consider the ideas that an individual may have in response to the inquiry shown in Figure 1. An individual can start by remembering the information about the definition of a magnetic force learned from textbooks as a lecture: "... Whether a magnet attracts or repels other magnets depends on the positions of its poles, or ends ... If different poles are closest to each other, the magnets attract ... If the same poles are closest to each other, the magnets repel ... ". Then, a chain of reasoning can be formed as an explanation: "The north pole

of one magnet is closest to the south pole of the other magnet. → Poles that are different attract. → These magnets will attract each other." This finally leads to the correct answer: "These magnets attract each other."

Early exploration of multimodal CoT involved transforming the inputs of different modalities into one modality and prompting LLMs to answer. One viable solution is to extract the caption of an image and concatenate it using the original language modality as input (Lu et al., 2022, 2023). However, a simplified caption cannot encompass all the details expressed by the image, leading to information loss in the reasoning process. Using only these captions may result in a lack of mutual synergy in the latent space of the multimodality. An alternative solution to facilitate the interaction between modalities is to fine-tune small models with cross-attentions to align multimodal features. Nevertheless, previous studies have shown that models trained with fewer than 100 billion parameters tend to produce illogical CoTs with hallucinated rationales (Ho et al., 2022; Magister et al., 2022). The challenge lies in the fact that the language model (LM) does not see images during pretraining and thus has no information about visual elements or methods to exploit vision features.

Recent studies have suggested incorporating both language and vision modalities into a two-stage framework, that is, rationale generation and answer inference (Zhang et al., 2023b). Instead of prompting the LM with an image caption, the vision features are extracted using a vision encoder and fed to a decoder along with the encoded language

---

\*Corresponding author

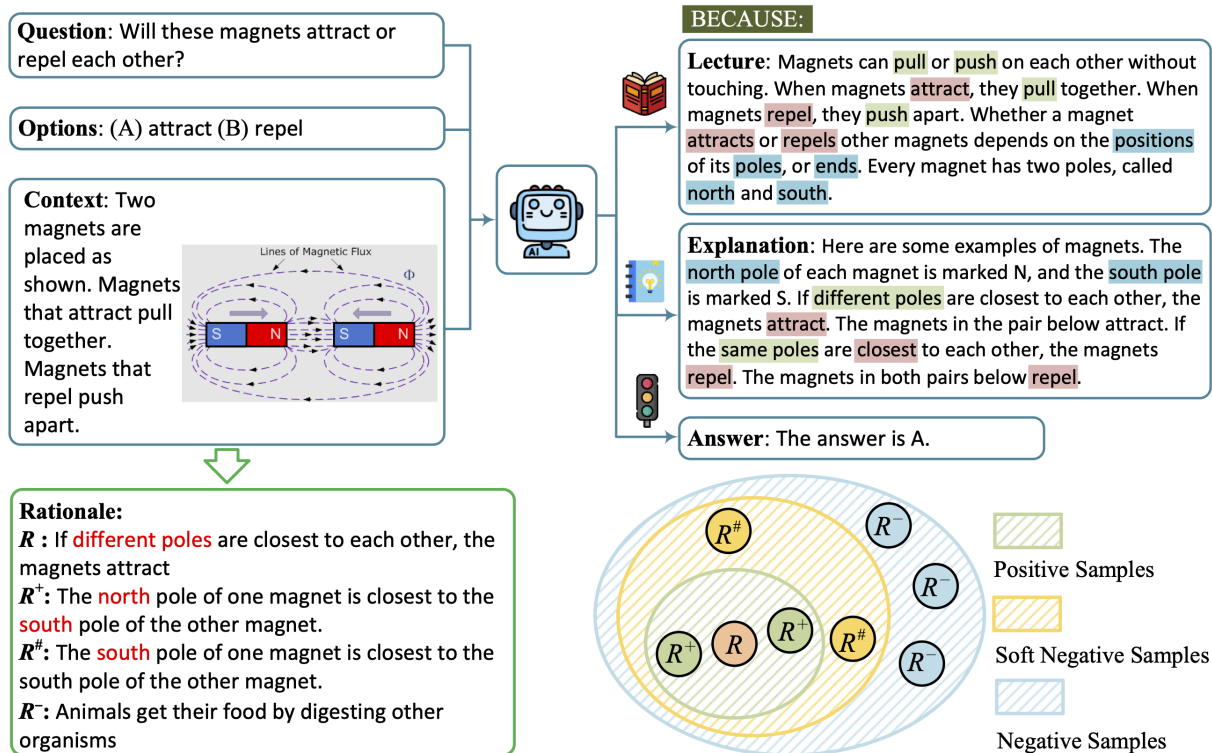


Figure 1: The latent distribution of the samples.  $R^+$  represents positive samples,  $R^\#$  represents soft negative samples, and  $R^-$  represents negative samples.

representation.

Although vision features are beneficial for better rationale generation, many reported errors stem from hallucinations (Zhang et al., 2023b). Considering the same example in Figure 1, an appropriate rationale is, *The north pole of one magnet is closest to the south of the other magnet*. However, simply modifying one word can make the rationale unreasonable, that is, *The south pole of one magnet is closest to the south of the other magnet*. For the decoder of generation, this inappropriate rationale can achieve an extremely low negative log-likelihood but will finally mislead the answer inference.

One viable solution to mitigate the hallucinated generation is to treat these inappropriate rationales as negative samples for contrastive learning. Negative samples are difficult to exclude if textual features are used to measure the distribution of the latent space. Several recent studies have defined these rationales as soft negative samples. Here, *negative* denotes that the samples differ semantically from the originals, whereas *soft* denotes that the samples share many textual similarities with the originals and cannot be simply regarded as pure negative samples.

This study proposes a rationale generation method using soft negative sampling (SNSE-CoT) to mitigate hallucinations in multimodal CoT. Contrastive learning was introduced to enhance rationale generation, and five methods were applied to

generate soft negative samples that shared highly similar text but had different semantics from the original. Bidirectional margin loss (BML) was applied to introduce them into the traditional contrastive learning framework that involves only positive and negative samples.

Extensive experiments were conducted on the ScienceQA dataset (Lu et al., 2022) to validate the effectiveness of the SNSE-CoT. The results showed that the proposed model outperformed models of previous studies in most categories for multimodal CoT.

The remainder of this paper is organized as follows. Section 2 reviews the preliminary knowledge. Section 3 describes the proposed SNSE-CoT in detail. Section 4 summarizes the experimental settings and empirical results. Section 5 briefly reviews the related works. Conclusions are drawn in Section 6.

## 2. Preliminary

A multimodal CoT applies two-stage training, including rationale generation and answer inference. Figure 2 shows the overall architecture of the two-stage model. An input  $X$  consists of an image input  $X_v$  and a language input  $X_l$ , where  $Q$  denotes the question text and  $C$  denotes the context text. The goal is to select correct answer  $A$  from multiple options  $M$ . To implement a multimodal CoT, the first

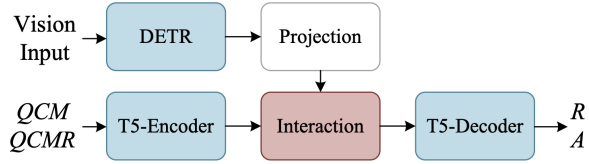


Figure 2: The overall architecture of the two-stage model.

stage is rationale generation, in which the model is required to generate a rationale.

$$R = f(X) \quad (1)$$

For answer inference, rationale  $R$  is appended to original language input  $\bar{X}_l$  as a new input, that is,  $\hat{X}_l = \bar{X}_l \circ R$ , where  $\circ$  is a concatenate operator. Then, updated input  $X' = \{\hat{X}_l, X_v\}$  is fed into the model to infer the final answer.

$$A = g(X') \quad (2)$$

In both stages, two independent transformer-based models  $f$  and  $g$  with the same architecture are trained. The aliases of both the stages can be represented as  $X \rightarrow R$  and  $X' \rightarrow A$ , respectively.

**Encoding.** For both stages, the vision and language inputs are  $X_v$  and  $X_l \in \{\bar{X}_l, \hat{X}_l\}$ , respectively, where  $\bar{X}_l$  is used for rationale generation, and  $\hat{X}_l$  is used for answer inference. The T5 encoder (Carion et al., 2020) is used to encode language input  $\hat{X}_l$ , and the DETR (Raffel et al., 2020) vectorizes vision input  $X_v$  into vision features.

$$H_l = \text{T5-Encoder}(X_l) \quad (3)$$

$$H_v = W_h \cdot \text{DETR}(X_v) \quad (4)$$

where  $H_l \in \mathbb{R}^{n \times d}$  is the hidden representation of the last layer of the T5 encoder,  $n$  denotes the length of the language input, and  $d$  denotes the dimensionality.  $H_v \in \mathbb{R}^{m \times d}$  is a vision feature, where  $m$  denotes the number of image patches, and  $W_h \in \mathbb{R}^{d \times d_v}$  is a linear projection used to transform the dimensionality from  $d_v$  to  $d$ .

**Interaction.** To integrate both vision and language representations, cross-attention with single-head self-attention is used to align text tokens with image tokens.

$$\hat{H}_v = \text{softmax} \left[ \frac{(W_Q H_l) \cdot (W_K H_v)^\top}{\sqrt{d}} \right] (W_V H_v) \quad (5)$$

where  $W_Q$ ,  $W_K$ , and  $W_V$  denote weight matrices for self-attention. Subsequently, a gated fusion mechanism is applied to integrate both features that is denoted as follows:

$$\sigma = \text{sigmoid}(W_l H_l + W_v \hat{H}_v) \quad (6)$$

$$H_{\text{Enc}} = (1 - \sigma) \cdot H_l + \sigma \cdot \hat{H}_v \quad (7)$$

where  $W_l$  and  $W_v$  denote trainable matrices.

**Decoding.** The model predicts the probability of generating a target  $Y \in \{R, A\}$  with length  $N$ . The models are trained by minimizing the negative likelihood loss.

$$\mathcal{L}_{\text{NLL}} = - \sum_{i=1}^N \log p_\theta(Y_i | X_l, X_v, Y_{<i}) \quad (8)$$

where  $\theta$  denotes all trainable parameters of either  $f$  or  $g$ .

### 3. Mitigating Hallucinated Generation

Both the quality and semantic correctness of rationale generation ultimately affect the choice of answer inference. This study proposes enhancing the ability to discriminate soft negative samples by introducing a bidirectional margin loss.

#### 3.1. Soft Negative Sampling

A high-quality soft negative sample is indistinguishable from the target sample and has different semantics for key information. Typically, high-quality soft negative samples are difficult to generate but can be obtained by modifying the target sample and observing the following three principles.

- **Principal 1:** *Soft negative samples have different semantics from the target sample for key information.* As shown in Figure 3(a), if this cannot be guaranteed, then the introduction of soft negative samples may move the generated samples away from the positive samples.
- **Principal 2:** *The modification of the target sample should be few.* Soft negative samples have high text similarity and a highly similar distribution with the target sample. The fewer the modifications to the target sample, the higher the text similarity and the higher the difficulty to distinguish the soft negative samples from the target sample. The number of modifications can be moderately increased according to the increase in the length of the target sample. As shown in Figure 3(b), if excessive modifications are introduced, the soft negative samples are far from the distribution of the positive samples, resulting in a negligible impact.
- **Principal 3:** *The generation of soft negative samples should be multiple and random.* Multiple soft negative samples should be distributed as evenly as possible around the positive sample region in the latent space. As shown in Figure 3(c), when the soft negative samples are on the same side as the target sample, they may push the generated samples away from the positive sample region instead of toward the target sample.

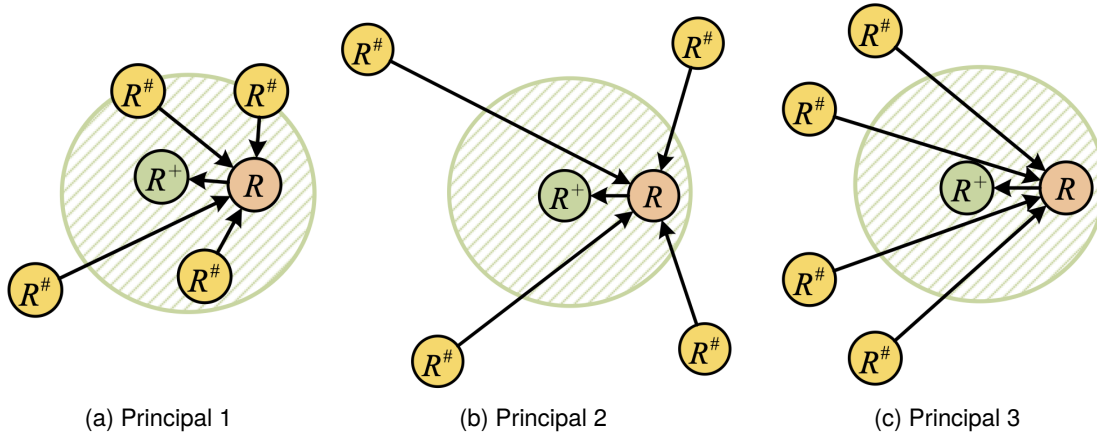


Figure 3: Possible situations arising from non-observance of the modification principle.  $R$  is the generated rationale.

The proposed soft negative sampling is specified using five methods as follows:

- **Affirmation-Negation Transformation.** Apply explicit negation with negative words, and based on the parsing information of SpaCy<sup>1</sup>, convert the sentences into syntactically correct and semantically clear negations.
- **Number Transformation.** Randomly select some numbers in the sentence and replace them with random numbers of equal length. When the length is greater than 1, ensure that the first number is not 0.
- **Orientation Transformation.** Randomly select some orientation words in the sentence and replace them with the opposite direction.
- **Unit Transformation.** Randomly select some unit words in the sentence and replace them with other units of the same category at random.
- **Option Transformation.** Randomly select some strings that contain the correct option in the sentence and replace them with other options.

Specifically, if other transformations fail to be implemented, we use an affirmation-negation transformation.

The modification of key information varies depending on the form of the sample. For multimodal reasoning, the explanation is modified first and then the lecture. The explanation provides reasoning ideas for a specific problem, which is considered to contain more important information, whereas the lecture typically provides methodological guidance, examples, and background knowledge for solving the problem. The rationale is a concatenation of lectures and explanations. If the rationale is devoid of any content, we use *Not* as a rationale.

<sup>1</sup><https://github.com/explosion/spaCy>.

### 3.2. Bidirectional Margin Loss

The generated positive and soft negative samples of the target rationale are further extracted using the following steps:

$$E_R = \text{emb}(R) \quad (9)$$

$$h = \text{mean}(W_R E_R + b_R) \quad (10)$$

where  $\text{emb}(\cdot)$  represents the embedding layer of the encoder in the rationale generation, and  $W_R$  and  $b_R$  represent the trainable matrix and bias, respectively.

Cosine similarity difference  $\Delta$  between positive and soft negative pairs is calculated as follows:

$$\Delta = \cos(h_i, h_{i_j}^\#) - \cos(h_i, h_i^+) \quad (11)$$

where  $\cos(\cdot)$  is the cosine similarity,  $h_i$  is the representation of the generated sample,  $h_i^+$  denotes the corresponding positive sample, and  $h_{i_j}^\#$  denotes the corresponding  $j$ -th soft negative sample. BML is used to model semantic similarity differences.

$$\mathcal{L}_{\text{BML}} = \frac{1}{k} \sum_{j=1}^k (\text{ReLU}(\Delta + \alpha) + \text{ReLU}(-\Delta - \beta)) \quad (12)$$

where  $k$  denotes the number of soft negative samples corresponding to each target sample, and  $\alpha$  and  $\beta$  denote the bottom and upper differences in semantic similarity difference between the positive and soft negative pairs, respectively. The BML aims to constrain  $\Delta$  within an interval of  $\Delta \in [-\beta, -\alpha]$ .

### 3.3. Training Objective

The training objectives for rationale generation and answer inference are respectively expressed as

follows:

$$\mathcal{L}_{\text{RG}} = \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_{\text{BML}} \quad (13)$$

$$\mathcal{L}_{\text{AI}} = \mathcal{L}_{\text{NLL}} \quad (14)$$

where  $\lambda$  is used to balance the two losses.

## 4. Experiments

### 4.1. Dataset

Empirical experiments were conducted using the ScienceQA benchmark (Lu et al., 2022), which is the first multimodal question answering dataset with a detailed CoT. Science QA features 26 topics, 127 categories, and 379 skills, covering a wide range of domains. The benchmark dataset was divided into training, validation, and test splits using 12,726, 4,241, and 4,241 examples, respectively. Eight question categories included were natural science, social science, language science, textual context, pictorial context, no context, grades 1–6, and grades 7–12. For the rationale generation stage, the optimal model was selected based on the ROUGE-L score. The predicted results were evaluated based on accuracy.

### 4.2. Implementation Details

UnifiedQA (Khashabi et al., 2020) was used to initialize the T5 model in two stages because it achieved the best fine-tuning results in the experiments by Lu et al. (2022). The model was fine-tuned for up to 20 epochs at a learning rate of 5e-5. The maximum input sequence lengths were 512 and 64 in the rationale generation and answer inference stages, respectively. For the soft negative samples, one sample was generated for each generation method in each round, and each sample was modified randomly in only one place.  $\alpha$ ,  $\beta$  and  $\lambda$  were set to 0.1, 0.3 and 0.1, respectively. These three parameters are further explored in Section 4.5. The random seed number was set to 42 to ensure reproducibility.

### 4.3. Baselines

For comparison, three categories of baseline models were selected as follows:

- MCAN (Yu et al., 2019), Top-Down (Anderson et al., 2018), BAN (Kim et al., 2018), DFAF (Gao et al., 2019), ViLT (Kim et al., 2021), Patch-TRM (Lu et al., 2021), and VisualBERT (Li et al., 2020).
- UnifiedQA<sub>Base</sub> w/CoT (Lu et al., 2022), Multimodal-CoT<sub>Base</sub>, and Multimodal-CoT<sub>Large</sub> (Zhang et al., 2023b).
- GPT-3.5 w/CoT (Lu et al., 2022), LLaMA-Adapter (Zhang et al., 2023a), LLaVa, LLaVa

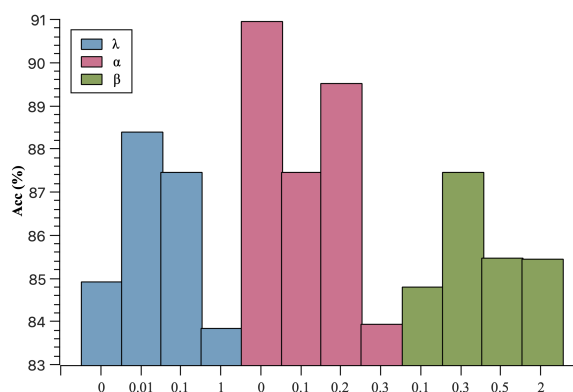


Figure 4: Hyperparameters fine-tuning.

(GPT-4) (Liu et al., 2023), and Chameleon (GPT-4) (Lu et al., 2023).

More details are presented in Appendix A.

### 4.4. Comparative Results

Table 1 summarizes the experimental results of the proposed method relative to those of the baselines. SNSE-CoT<sub>Base</sub> performed similarly to humans. SNSE-CoT<sub>Large</sub> outperformed all the previous methods to achieve the current best performance. Compared with Multimodal-CoT, the average performance of SNSE-CoT increased by approximately 2.5 to 3%. Moreover, SNSE-CoT<sub>Large</sub> was worse than LLaVa (GPT-4) only for social science problems (SOC) for all types of problems, indicating that the soft negative sample generation method designed in this study improved the correctness of the model for generating various types of CoTs. In particular, SNSE-CoT<sub>Large</sub> achieved substantial improvement in problems with paired images (IMG), becoming the first method to exceed 90% performance on this type of problem. This indicated that the contrastive learning approach strengthened the model’s ability to accurately understand images of specific problems.

### 4.5. Hyperparameter Fine-Tuning

To explore the effect of hyperparameters  $\lambda$  in Eq. (13) and  $\alpha$  and  $\beta$  in Eq. (12), a grid strategy was used. For balanced parameter  $\lambda$ , candidate set  $\{0, 0.01, 0.1, 1\}$  was used; for bottom difference  $\alpha$  of the cosine similarity difference, the candidate set  $\{0, 0.1, 0.2, 0.3\}$  was used; and for upper difference  $\beta$  of the cosine similarity difference, candidate set  $\{0.1, 0.3, 0.5, 2\}$  was used.  $\beta = 2$  is a special value indicating that the upper difference is ignored because the range of the cosine similarity is -1.0 to 1.0. The results of the parametric analysis are shown in Figure 4.

Balance parameter  $\lambda$  affects the strength of the soft negative sample rejection. If  $\lambda$  is extremely

Model	Size	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Human	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
MCAN	95M	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72	54.54
Top-Down	70M	59.50	54.33	61.82	62.90	54.88	59.79	57.27	62.16	59.02
BAN	112M	60.88	46.57	66.64	62.61	52.60	65.51	56.83	63.94	59.37
DFAF	74M	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA <sub>Base</sub> w/CoT	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
GPT-3.5 w/CoT	175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
Multimodal-CoT <sub>Base</sub>	223M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
Multimodal-CoT <sub>Large</sub>	738M	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
LLaMA-Adapter	6B	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LLaVa	13B	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVa (GPT-4)	13B	91.56	<b>96.74</b>	91.09	90.62	88.99	93.52	92.73	92.16	92.53
Chameleon (GPT-4)	-	89.83	74.13	89.82	88.27	77.64	92.13	88.03	83.72	86.54
SNSE-CoT <sub>Base</sub>	223M	90.05	78.85	89.09	89.64	84.78	90.38	87.67	87.08	87.46
SNSE-CoT <sub>Large</sub>	738M	<b>96.80</b>	90.33	<b>93.09</b>	<b>96.73</b>	<b>93.36</b>	<b>94.08</b>	<b>94.71</b>	<b>94.07</b>	<b>94.48</b>

Table 1: Comparison of the experiment results (%). Size = backbone model size. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Part 1: Human performance; Part 2: Traditional VQA; Part 3: Small model with CoT; Part 4: Large model with CoT; Part 5: Our SNSE-CoT results. Results in bold are the best performance.

small, soft negative samples play a limited role and performance grows less. When  $\lambda$  is extremely large, the generated samples have difficulty in clustering toward the target center, and the performance may even become worse.  $\alpha$  and  $\beta$  constrain the range of cosine similarity difference, and an appropriate bottom and upper difference can play a better role in soft negative samples. The effect of bottom difference  $\alpha$  on performance is more important.

#### 4.6. Ablation Studies

Table 2 reports the results of the ablation experiments. To investigate the effectiveness of the proposed soft negative sample generation method, the number, orientation, unit, and option transformations were removed separately. For the number transformation, a slight decrease in performance was observed after removal, indicating that the method did not considerably enhance the model’s numerical understanding. For the orientation and unit transformations, which further enhance the model’s understanding of the map and its ability to compare values, a decrease in performance of approximately 1% was observed after removal. For the option transformation, a significant decrease in model performance was observed after removal because it is a general transformation that can modify keywords well for QA problems.

The principle of soft negative sample generation was also analyzed. The randomness of **Principle 3** was ablated by fixing the soft negative sam-

Model	Avg
SNSE-CoT <sub>Base</sub>	87.46
w/o number	87.12
w/o orientation	86.23
w/o unit	86.03
w/o option	85.69
w/o random	85.57

Table 2: Ablation study on SNSE-CoT (%).

Number	1	2	3	all
Changes	0	-0.59	+0.08	+0.17

Table 3: The impact of the amount of modification on performance (%). All means all modifications.

ples for each epoch. As observed from the results, the ablation of randomness did not guarantee that the soft negative samples were evenly distributed around the positive sample area, and the performance yielded a significant degradation of approximately 2%.

To explore the effect of the number of modifications mentioned in **Principle 2**, the relevant experimental results are reported in Table 3. The small effect of the number of modifications on the performance might be due to the small number of modifiable positions in each CoT and the limited increase in the distance of the soft negative samples from the target center as the number of modifications

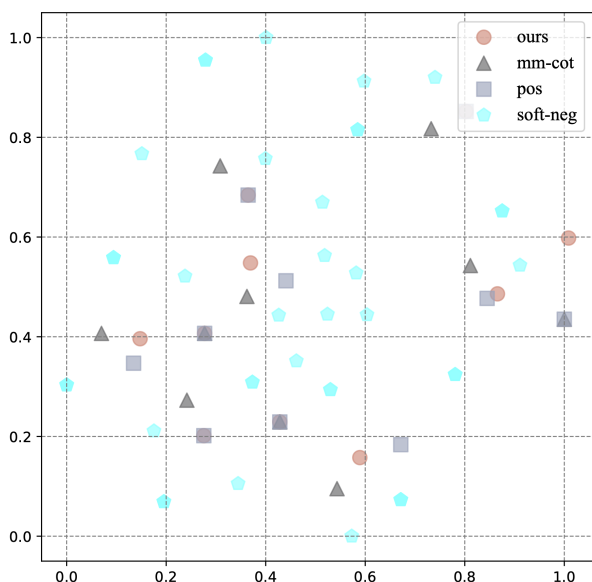


Figure 5: Visual latent distribution. “ours” represents samples generated by SNSE-CoT, “mm-cot” represents samples generated by Multimodal-CoT, “pos” represents positive samples, and “soft-neg” represents soft negative samples.

increased.

#### 4.7. Visual Latent Distribution of Samples

We randomly selected ten samples, and the samples generated by SNSE-CoT, samples generated by Multimodal-CoT, positive samples, and soft negative samples of each sample were further extracted as feature  $h$  using Eq. (9) and (10): These functions were visualized using the tSNE tool.

Figure 5 illustrates this visualization. The soft negative samples were largely evenly distributed around the positive samples. This was consistent with the expectations of this study, and demonstrated the high quality of the soft negative samples generated by the proposed method. However, soft negative samples were still distributed on one side or far from each other, and our method must be improved. In terms of the distribution distance between the samples generated by Multimodal-CoT, the samples generated by SNSE-CoT, and the target positive samples, the distribution of the samples generated by SNSE-CoT was generally closer to the target samples, or even overlapped. However, worse cases existed, in which the samples generated by SNSE-CoT were pushed far away from the positive sample area by the soft negative samples, suggesting that the generation of soft negative samples should more carefully follow the generation principles proposed in this study.

#### 4.8. Case Analysis

Several examples were selected to further illustrate the effectiveness of SNSE-CoT in improving the quality of CoT generation. Table 4 lists the selected examples corresponding to golden CoT, CoT generated by Multimodal-CoT, and CoT generated by SNSE-CoT. A more detailed case analysis is provided in Appendix C. The results indicated that the proposed SNSE-CoT enhanced the map numerical aspects, affirmation-negation comprehension, and comparison of values in different units. In particular, the SNSE-CoT was still deficient in numerical comprehension and was only partially correct in the relevant example that corresponded to the results of the ablation experiment.

### 5. Related Work

#### 5.1. Traditional VQA

VQA is a series of tasks that provides a picture and natural language question related to that picture, and the computer can produce the correct answer. Since the VQA task was first proposed (Antol et al., 2015), many VQA datasets (Goyal et al., 2017; Hudson and Manning, 2019; Johnson et al., 2017) have been constructed to contribute to this research effort.

Researchers have proposed various approaches to improve the accuracy and interpretability of the models. Among them, joint embedding approaches (Li et al., 2020) jointly encode the image and question and then decode the features of the mixed modality to generate the answer. More studies have focused on the application of attention mechanisms (Anderson et al., 2018; Gao et al., 2019; Kim et al., 2018; Yuan et al., 2023) that have shown that models are more effective at capturing key image parts based on questions. Compositional models (Andreas et al., 2016; Xiong et al., 2016) provide a method to dynamically compose modules to generate answers based on the type of question. Other researchers have focused on introducing external knowledge bases (Wu et al., 2016) to supplement the knowledge required to answer questions by retrieving knowledge bases. Each of these models has its own characteristics; however, they are all black-box models that output only answers.

ScienceQA datasets containing multimodal contexts and different topics in the scientific domain have been proposed. Most answers are annotated with lectures and explanations that allow multimodal CoT reasoning to be elicited and the reasoning process to be explicitly presented.

ID	Golden CoT	MultiModal-CoT	SNSE-CoT
6902	... This is the <b>Pacific Ocean</b> .	... This is the <b>Indian Ocean</b> .	... This is the <b>Pacific Ocean</b> .
7301	... there were <b>6</b> solute particles on the left side of the membrane and <b>2</b> solute particles on ...	... there were <b>5</b> solute particles on the left side of the membrane and <b>6</b> solute particles on ...	... there were <b>5</b> solute particles on the left side of the membrane and <b>2</b> solute particles on ...
10130	... Animals <b>get</b> their food by digesting other organisms.	... Animals <b>cannot</b> their food by digesting other organisms.	... Animals <b>get</b> their food by digesting other organisms.
12785	... the temperature of the air on a hot day is <b>36°C</b> . 36°F is too <b>cold</b> .	... the temperature of the air on a hot day is <b>36°F</b> . 36°F is too <b>hot</b> .	... the temperature of the air on a hot day is <b>36°C</b> . 36°F is too <b>cold</b> .

Table 4: CoT outputs of different models. ID indicates the identification number of the question in the ScienceQA dataset. Keywords are labeled with different colors, where red refers to wrong words, green refers to correct words.

## 5.2. CoT Reasoning

CoT encourages LLMs to generate intermediate chains of reasoning to solve problems, and LLMs typically use two technical paradigms for CoT reasoning: zero-shot CoT (Kojima et al., 2022) and few-shot CoT (Wei et al., 2022; Zhang et al., 2022). Few-shot CoT uses step-by-step reasoning demonstrations as the conditions for reasoning, each of which contains the question and chain of reasoning leading to the final answer and can be produced manually or automatically, called Manual-CoT (Kojima et al., 2022) and Auto-CoT (Zhang et al., 2022), respectively. Effective demonstrations make few-shot CoT a stronger performer than zero-shot CoT and have attracted more research interest.

Certain studies have focused on the ability to inspire CoT reasoning in small models. However, models with 100 billion parameters tend to produce illogical CoTs, leading to incorrect answers (Wei et al., 2022). The performance substantially drops when small models are directly fine-tuned to generate CoTs to reason answers (Lu et al., 2022). Knowledge distillation is employed by fine-tuning the student model on the output of the CoT generated by the larger teacher model (Ho et al., 2022; Magister et al., 2022), and significant performance improvement is obtained.

As stated in Section 1, several studies focused on multimodal CoT reasoning. The critical challenge is to unify vision and language modalities. The images are converted into captions to prompt the LLMs for CoT inference (Lu et al., 2022, 2023). However, caption conversion loses considerable information, and researchers have attempted to interactively combine vision and language modalities in large and small models (Liu et al., 2023; Zhang et al., 2023a,b) to obtain remarkable results. For instance, Zhang et al. (2023b) indicated that the introduction of the visual modality supplemented more information and alleviated the hallucinations

of the small model CoT, and the small model CoT reasoning ability made a qualitative leap.

## 5.3. Contrastive Learning

Contrastive learning allows the models to learn from both positive and negative samples. Three key issues in contrastive learning are construction of positive and negative samples, design of the encoder, and selection of the loss function. The development of contrastive learning can be divided into four phases.

In the first phase, the various methods and models are not unified, nor are the objective functions and agent tasks (Wu et al., 2018; Ye et al., 2019). In the second phase (Chen et al., 2020a,b), the details tend to be uniform, objective function is InfoNCE or similar, and model is a combination of the encoder and projection head. Stronger data augmentation is used. Momentum encoders have been proposed to solve feature inconsistency problems. In the third phase, contrastive learning eliminates the use of negative samples (Chen et al., 2020b; Grill et al., 2020); this stage is a summary generalization of all methods. In the fourth phase, most studies have focused on the use of contrastive learning in transformers (Caron et al., 2021).

Recently, the concept of soft negative samples (Wang et al., 2022) was proposed to guide models to focus on semantic similarity and alleviate feature suppression.

Most negative samples for these methods originate from other samples within the same batch, and few studies have focused on the manner in which negative samples are generated.

## 6. Conclusions

This study proposed mitigating hallucinated rationale generation by using soft negative sampling



for multimodal scientific quiz questions to generate more accurate CoTs. Specifically, we designed five high-quality soft negative sample generation methods: affirmation-negation transformation, number transformation, orientation transformation, unit transformation, and option transformation. Bidirectional margin loss was used to enable the model to distinguish between soft negative samples. Experimental results showed that the proposed method outperformed the methods in the previous studies on the ScienceQA benchmark dataset and validated the effectiveness of the proposed methods.

Future work will attempt to design a general method for automatically generating soft negative examples such that the model self-corrects for small semantic differences.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos.61966038 and 62266051, and the Exam-Exempted Postgraduate Research and Innovation Foundation of Yunnan University under Grant No.TM-23236806. The authors would like to thank the anonymous reviewers for their constructive comments.

## Bibliographical References

- Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2022. Cont: Contrastive neural text generation. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS-2022)*.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2018)*, pages 6077–6086.
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2016)*, pages 39–48.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV-2015)*, pages 2425–2433.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the 16th European Conference on Computer Vision (ECCV-2020)*, pages 213–229.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV-2021)*, pages 9650–9660.
- Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020a. Big self-supervised models are strong semi-supervised learners. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS-2020)*, 33:22243–22255.
- Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2021)*, pages 15750–15758.
- Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C H Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In *Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2019)*, pages 6639–6648.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP-2021)*, pages 6894–6910.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2017)*, pages 6904–6913.
- Jean-Bastien Grill, Florian Strub, Florent Alché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi

- Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS-2020)*, volume 33, pages 21271–21284.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2019)*, pages 6700–6709.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1–38.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2017)*, pages 2901–2910.
- Daniel Khoshdel, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907.
- Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS-2018)*, pages 1571–1581.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML-2021)*, pages 5583–5594.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS-2022)*.
- Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL-2022)*, pages 6327–6337.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)*, pages 5265–5275.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS-2022)*.
- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.
- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS-2021)*.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.
- Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 244–258.
- Colin Raffel, Noam 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. *The Journal of Machine Learning Research*, 21:5485–5551.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples. *arXiv preprint arXiv:2201.05979*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS-2022)*, volume 35, pages 24824–24837.
- Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2016)*, pages 4622–4630.
- Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2018)*, pages 3733–3742.
- Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on Machine Learning (ICML-2016)*, pages 2397–2406.
- Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2019)*, pages 6210–6219.
- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-2019)*, pages 6281–6290.
- Li Yuan, Yi Cai, Jin Wang, and Qing Li. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the 37th AAAI conference on artificial intelligence (AAAI-2023)*, 9, pages 11051–11059.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR-2020)*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

## A. Baseline Models

For comparison, three categories of baseline models are selected as follows:

### A.1. Traditional VQA

Traditional VQA baselines consider the question, context, and choices as the textual input and the image as the vision input, and predict the score distribution over choice candidates via a linear classifier.

- MCAN (Yu et al., 2019) designs self-attention units and guided attention units, and constructs a new deep co-attention network through the combination and stacking between them.
- Top-Down (Anderson et al., 2018) proposes a new visual attention mechanism that combines bottom-up and top-down to allow attention to be calculated more naturally at the object and other prominent area level.
- BAN (Kim et al., 2018) uses variants of multimodal residual networks for joint representation and finally classification by MLP to predict answers.
- DFAF (Gao et al., 2019) proposes a multimodal feature fusion method using external and internal modal information flow.
- ViLT (Kim et al., 2021) and Patch-TRM (Lu et al., 2021) parse the diagram in a pyramid layout and apply cross-modal transformers with attention mechanism to learn the meaningful joint diagram-question feature.
- VisualBERT (Li et al., 2020) is a visual language pretraining model using a self-attention

mechanism to mine the relationship between text and image regions in the input image.

### A.2. Small Model with CoT

- UnifiedQA<sub>Base</sub> w/CoT (Lu et al., 2022) extracts the caption of an image and concatenates it with the original language modality as the input to fine-tune the LM.
- Multimodal-CoT<sub>Base</sub> and Multimodal-CoT<sub>Large</sub> (Zhang et al., 2023b) use gated fusion mechanisms to combine language and visual modalities into a two-stage framework.

### A.3. Large Model with CoT

- GPT-3.5 w/CoT (Lu et al., 2022) uses the same method as that of UnifiedQA<sub>Base</sub> w/CoT to prompt LLMs.
- LLaMA-Adapter (Zhang et al., 2023a) adjusts LLaMA (Touvron et al., 2023) to an instruction-following model by fine-tuning the adapter to insert vision features into the LM.
- LLaVa (Liu et al., 2023) is a multimodal large model fine-tuned using a multimodal instruction dataset.
- LLaVa (GPT-4) (Liu et al., 2023) collaborates with GPT-4 to first explain the cause and then infer the answer.
- Chameleon (GPT-4) (Lu et al., 2023) uses GPT-4 as a natural language planner to break down problems into chains of multiple tool combinations (design workflow) and then invoke tools to collaboratively solve problems.

## B. Role of Vision Features

To explore the impact of the proposed method on the LM, the vision features were further removed. Table 5 shows the differences in performance before and after the removal of vision features for different sizes of LLMs.

Removing the vision features significantly reduced the performances (approximately 12% and 9%, respectively) of both LLMs of different sizes, where the performance of the large model reduced relatively slightly. Large performance degradation was mainly generated in the IMG and image-related data. A noticeable performance degradation was also observed in non-image-related data because with the vision features removed, more error CoTs were generated in the rationale generation phase, further disrupting the inference logic in the answer inference phase.

## C. Examples of Case Studies

This section presents a detailed case study. As shown in Figures 6 and 7, the proposed method was well enhanced in terms of identifying the map regions. As shown in Figures 8 and 9, the unit comparison capability of the model also improved. However, Figures 10 and 11 show that a partially correct CoT may still lead to incorrect answer inferences, and this misdirection is more likely to occur when the conclusion sentence is incorrect. Similarly, as shown in Figures 12 and 13, guaranteeing the overall correctness of the CoT in terms of number understanding is difficult, even if some numbers can be improved. For common sense problems, models could easily learn common sense that required only memorization (see Figure 14) but struggled to learn common sense that required logical computation (see Figure 15).

Model	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
SNSE-CoT <sub>Base</sub>	90.05	78.85	89.09	89.64	84.78	90.38	87.67	87.08	87.46
w/o vision features	75.63	70.76	79.82	74.18	68.51	83.15	75.83	73.66	75.41
SNSE-CoT <sub>Large</sub>	96.80	90.33	93.09	96.73	93.36	94.08	94.71	94.07	94.48
w/o vision features	82.09	82.34	89.42	88.60	81.35	90.10	84.69	83.31	85.84

Table 5: Ablation results of vision features(%).

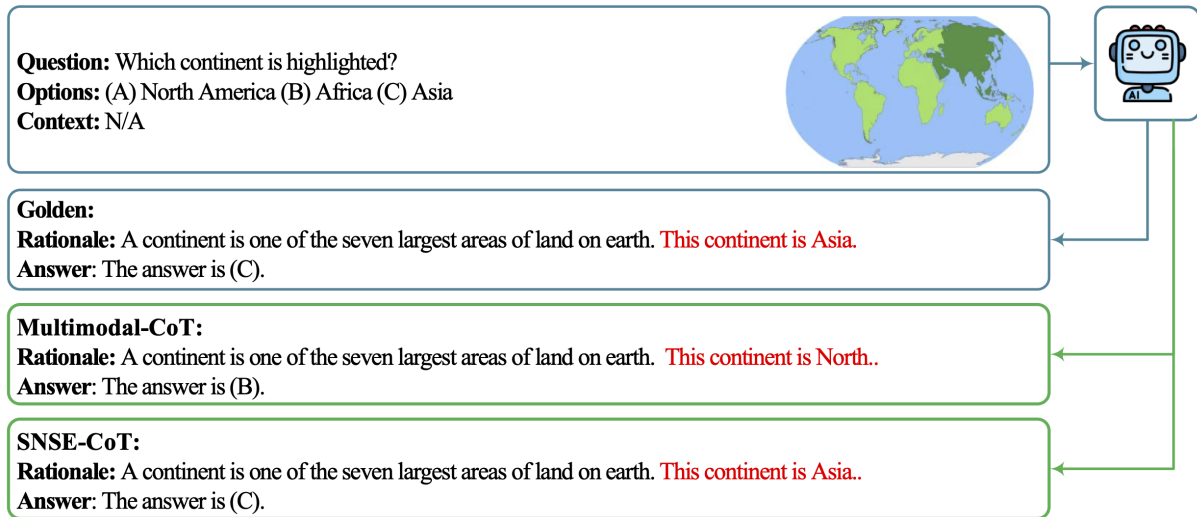


Figure 6: Example of ID 517.

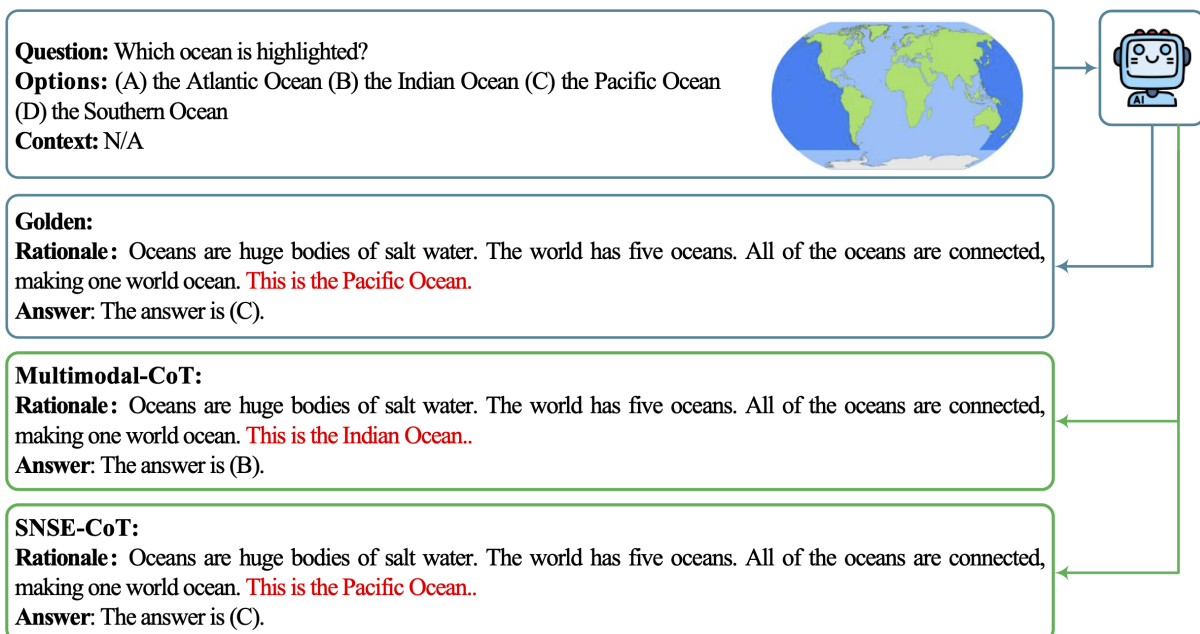



Figure 7: Example of ID 6902.

**Question:** What is the temperature of the air on a warm, sunny day?  
**Options:** (A) 26°C (B) 26°F  
**Context:** Select the better estimate.

N/A



**Golden:**  
**Rationale:** Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means ... The better estimate for the temperature of the air on a warm, sunny day is 26°C. 26°F is too cold.  
**Answer:** The answer is (A).


**Multimodal-CoT:**  
**Rationale:** Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means ... The better estimate for the temperature of the air on a warm, sunny day is 26 °F. 26 °C is too cold.  
**Answer:** The answer is (B).

**SNSE-CoT:**  
**Rationale:** Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means ... The better estimate for the temperature of the air on a warm, sunny day is 26°C. 26 °F is too cold.  
**Answer:** The answer is (B).

Figure 8: Example of ID 348.

**Question:** What is the temperature of the air on a hot day?  
**Options:** (A) 36°F (B) 36°C  
**Context:** Select the better estimate.

N/A



**Golden:**  
**Rationale:** Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means ... The better estimate for the temperature of the air on a hot day is 36°C. 36°F is too cold.  
**Answer:** The answer is (B).

**Multimodal-CoT:**  
**Rationale:** Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means ... The better estimate for the temperature of the air on a hot day is 36°F. 36°C is too hot.  
**Answer:** The answer is (A).

**SNSE-CoT:**  
**Rationale:** Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means ... The better estimate for the temperature of the air on a hot day is 36°C. 36°F is too cold.  
**Answer:** The answer is (B).

Figure 9: Example of ID 12785.

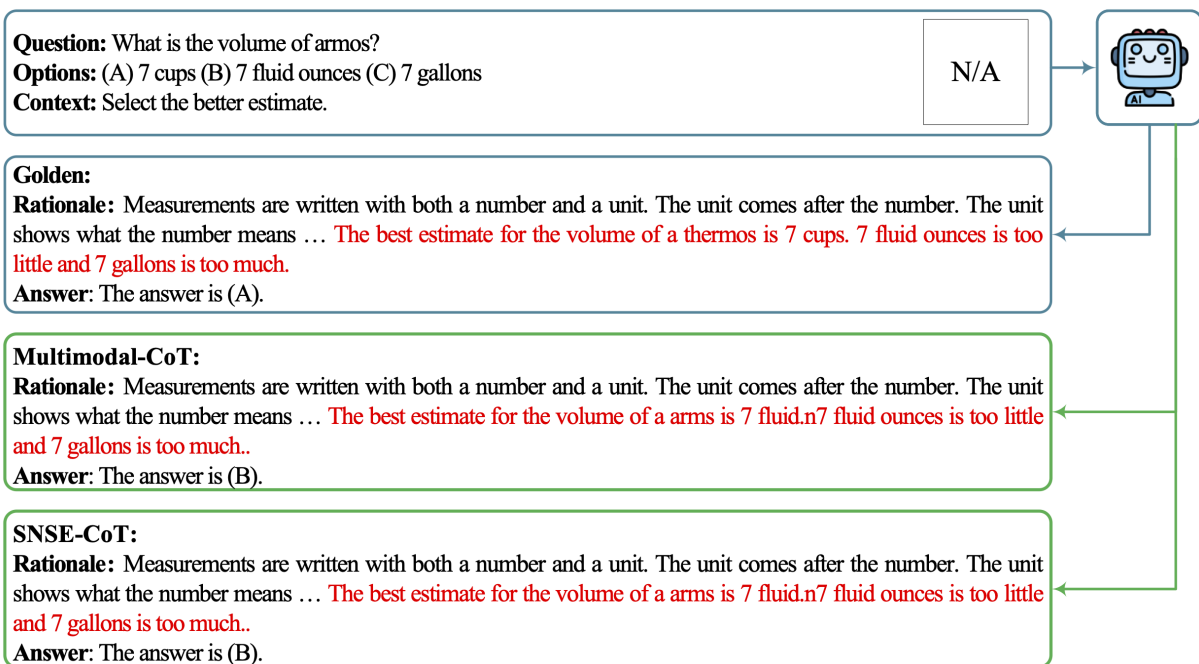


Figure 10: Example of ID 9653.

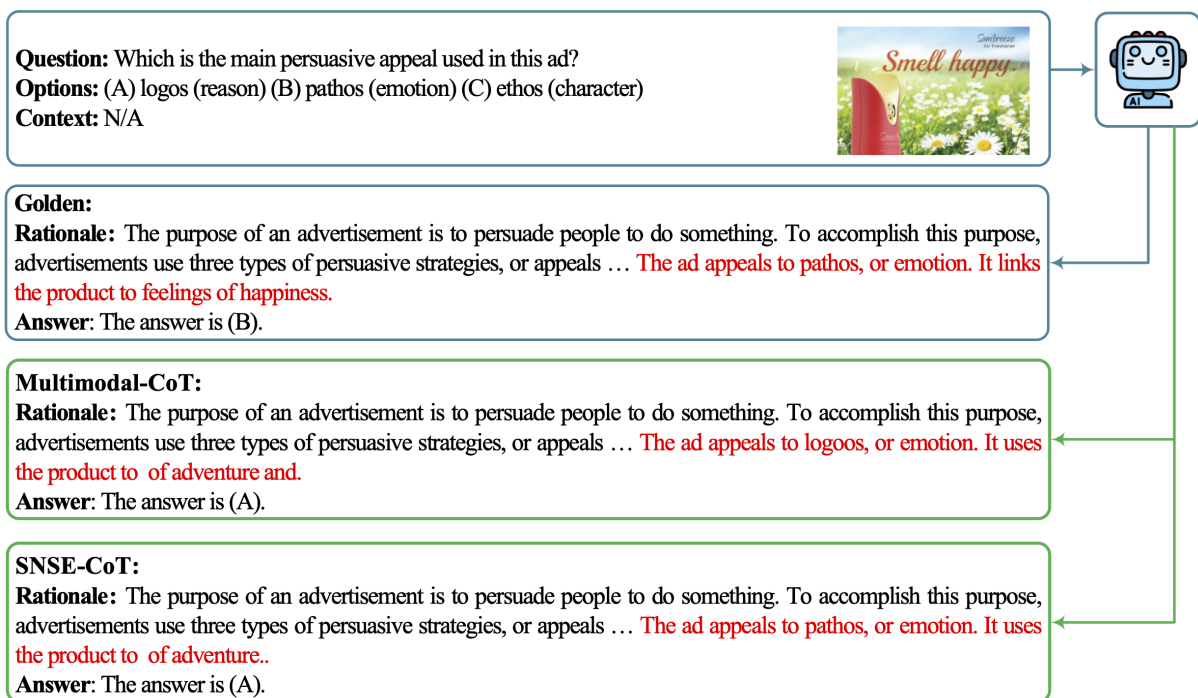
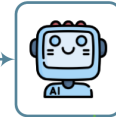
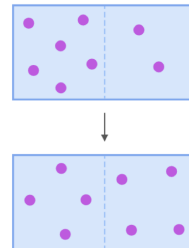


Figure 11: Example of ID 6964.

**Question:** Complete the text to describe the diagram. Solute particles moved in both directions across the permeable membrane. But more solute particles moved across the membrane (). When there was an equal concentration on both sides, the particles reached equilibrium.

**Options:** (A) to the right than to the left (B) to the left than to the right

**Context:** The diagram below shows a solution with one solute. Each solute particle is represented by a pink ball. The solution fills a closed container that is divided in half by a membrane. The membrane, represented by a dotted line, is permeable to the solute particles. The diagram shows how the solution can change over time during the process of diffusion.



**Golden:**

**Rationale:** In a solution, solute particles move and spread throughout the solvent. The diagram below shows how a solution can change over time. Solute particles move from the area where they are at a higher concentration to the area where they are at a lower concentration ... Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 6 solute particles on the left side of the membrane and 2 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 4 solute particles on each side of the membrane. There were 2 more solute particles on the right side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the right than to the left.

**Answer:** The answer is (A).

**Multimodal-CoT:**

**Rationale:** In a solution, solute particles move and spread throughout the solvent. The diagram below shows how a solution can change over time. Solute particles move from the area where they are at a higher concentration to the area where they are at a lower concentration ... Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 5 solute particles on the left side of the membrane and 6 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 5 solute particles on each side of the membrane. There was 2 more solute particles on the right side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the right than to the left..

**Answer:** The answer is (A).

**SNSE-CoT:**

**Rationale:** In a solution, solute particles move and spread throughout the solvent. The diagram below shows how a solution can change over time. Solute particles move from the area where they are at a higher concentration to the area where they are at a lower concentration ... Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 6 solute particles on the left side of the membrane and 2 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 5 solute particles on each side of the membrane. There was 2 more solute particles on the right side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the right than to the left..

**Answer:** The answer is (A).

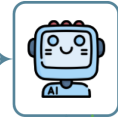
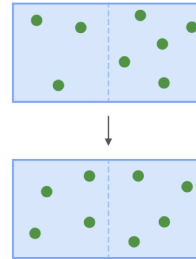
Figure 12: Example of ID 7301.



**Question:** Complete the text to describe the diagram. Solute particles moved in both directions across the permeable membrane. But more solute particles moved across the membrane (). When there was an equal concentration on both sides, the particles reached equilibrium.

**Options:** (A) to the left than to the right (B) to the right than to the left

**Context:** The diagram below shows a solution with one solute. Each solute particle is represented by a green ball. The solution fills a closed container that is divided in half by a membrane. The membrane, represented by a dotted line, is permeable to the solute particles. The diagram shows how the solution can change over time during the process of diffusion.



**Golden:**

**Rationale:** In a solution, solute particles move and spread throughout the solvent. The diagram below shows how a solution can change over time. Solute particles move from the area where they are at a higher concentration to the area where they are at a lower concentration ... Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 3 solute particles on the left side of the membrane and 5 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 4 solute particles on each side of the membrane. There was 1 more solute particle on the left side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the left than to the right.

**Answer:** The answer is (A).

**Multimodal-CoT:**

**Rationale:** In a solution, solute particles move and spread throughout the solvent. The diagram below shows how a solution can change over time. Solute particles move from the area where they are at a higher concentration to the area where they are at a lower concentration ... Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 5 solute particles on the left side of the membrane and 4 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 5 solute particles on each side of the membrane. There was 1 more solute particle on the right side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the right than to the left.

**Answer:** The answer is (B).



**SNSE-CoT:**

**Rationale:** In a solution, solute particles move and spread throughout the solvent. The diagram below shows how a solution can change over time. Solute particles move from the area where they are at a higher concentration to the area where they are at a lower concentration ... Look at the diagram again. It shows you how the solution changed during the process of diffusion. Before the solute particles reached equilibrium, there were 5 solute particles on the left side of the membrane and 4 solute particles on the right side of the membrane. When the solute particles reached equilibrium, there were 5 solute particles on each side of the membrane. There was 1 more solute particle on the right side of the membrane than before. So, for the solute particles to reach equilibrium, more solute particles must have moved across the membrane to the right than to the left.

**Answer:** The answer is (B).

Figure 13: Example of ID 4740.

**Question:** Can *Leopardus wiedii* cells make their own food?  
**Options:** (A) no (B) yes  
**Context:** This organism is *Leopardus wiedii*. It is a member of the animal kingdom. *Leopardus wiedii* is commonly called a margay. Margays live in the forest in Central and South America. Margays are nocturnal. A nocturnal animal sleeps most of the day and is awake at night.



**Golden:**  
**Rationale:** In the past, scientists classified living organisms into two groups: plants and animals ... *Leopardus wiedii* is an animal. Animal cells cannot make their own food. **Animals get their food by digesting other organisms.**  
**Answer:** The answer is (A).

**Multimodal-CoT:**  
**Rationale:** In the past, scientists classified living organisms into two groups: plants and animals ... *Leopardus wiedii* is an animal. Animal cells cannot make their own food. **Animals cannot their food by digesting other organisms.**  
**Answer:** The answer is (B).

**SNSE-CoT:**  
**Rationale:** In the past, scientists classified living organisms into two groups: plants and animals ... *Leopardus wiedii* is an animal. Animal cells cannot make their own food. **Animals get their food by digesting other organisms.**  
**Answer:** The answer is (A).

Figure 14: Example of ID 10130.

**Question:** Which word would you find on a dictionary page with the following guide words?helping - hunter  
**Options:** (A) hospital (B) hardship  
**Context:** N/A

**Golden:**  
**Rationale:** Guide words appear on each page of a dictionary. They tell you the first word and last word on the page ... **Put the words in alphabetical order.**Since hospital is between the guide words helping - hunter, it would be found on that page.  
**Answer:** The answer is (A).

**Multimodal-CoT:**  
**Rationale:** Guide words appear on each page of a dictionary. They tell you the first word and last word on the page ... **Put the words in alphabetical order.**Since hardship is between the guide words helping - hunter, it would be found on that page..  
**Answer:** The answer is (B).

**SNSE-CoT:**  
**Rationale:** Guide words appear on each page of a dictionary. They tell you the first word and last word on the page ... **Put the words in alphabetical order.**Since hardship is between the guide words helping - hunter, it would be found on that page..  
**Answer:** The answer is (B).

Figure 15: Example of ID 13068.