Multi-Modal Latent Space Learning for Chain-of-Thought Reasoning in Language Models

Liqi He[1,†], **Zuchao Li**[1,†*], **Xiantao Cai**[1], **Ping Wang**[2,3]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, 430072, China
[2]Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China
[3]School of Information Management, Wuhan University, Wuhan 430072, China
heliqi@whu.edu.cn, zcli-charlie@whu.edu.cn, caixiantao@whu.edu.cn, wangping@whu.edu.cn

Abstract

Chain-of-thought (CoT) reasoning has exhibited impressive performance in language models for solving complex tasks and answering questions. However, many real-world questions require multi-modal information, such as text and images. Previous research on multi-modal CoT has primarily focused on extracting fixed image features from off-the-shelf vision models and then fusing them with text using attention mechanisms. This approach has limitations because these vision models were not designed for complex reasoning tasks and do not align well with language thoughts. To overcome this limitation, we introduce a novel approach for multi-modal CoT reasoning that utilizes latent space learning via diffusion processes to generate effective image features that align with language thoughts. Our method fuses image features and text representations at a deep level and improves the complex reasoning ability of multi-modal CoT. We demonstrate the efficacy of our proposed method on multi-modal ScienceQA and machine translation benchmarks, achieving state-of-the-art performance on ScienceQA. Overall, our approach offers a more robust and effective solution for multi-modal reasoning in language models, enhancing their ability to tackle complex real-world problems.

Introduction

In our daily lives, we are constantly bombarded with information from various sources, such as text, images, and more. To make sense of this complex world, we need to be able to acquire and integrate multi-modal information effectively. For example, as shown in Figure 1, when we see the slogan "Please keep off the grass" (language modality) on the lawn of the park and a child is playing football on the same lawn (visual modality) , we think of the negative effects of trampling the grass on the park's ecological environment and prepare to told the child to go to the football field (language thought) . These ideas come from our deep understanding and reasoning of linguistic and visual information, which can be called language thought.

Figure 1: Language Thought

In recent years, chain-of-thought (CoT), which involves a series of intermediate reasoning steps (also known as rationale), has significantly enhanced the complex reasoning ability of large language models by providing them with access to a portion of language thought (Wei et al. 2022).

Training smaller language-only models with less than 100 billion parameters for CoT reasoning remains a significant challenge due to hallucination and tend to produce illogical rationales. To address these problems more effectively, it is crucial to enable large language models to develop a deeper understanding of multi-modal information and generate more effective language thought. One solution that has been proposed to help integrate information across visual and linguistic modalities is Multi-Modal CoT (MM-CoT) (Zhang et al. 2023b). MM-CoT extracts fixed image features and text representations and fused them to obtain multi-modal features. MM-CoT adopts a two-stage framework that includes rationale generation and answer inference, as shown in Figure 2 (c). This approach has been shown to outperform generating rationale and answer together on question answering tasks (Zhang et al. 2023b).

However, existing multi-modal CoT models rely on fixed image features extracted by pre-trained vision extraction

* Corresponding author. † Equal contribution. This work was supported by the National Natural Science Foundation of China (No. 62306216), the Natural Science Foundation of Hubei Province of China (No. 2023AFB816), the Fundamental Research Funds for the Central Universities (No. 2042023kf0133), National Natural Science Foundation of China [No. 72074171] [No. 72374161].

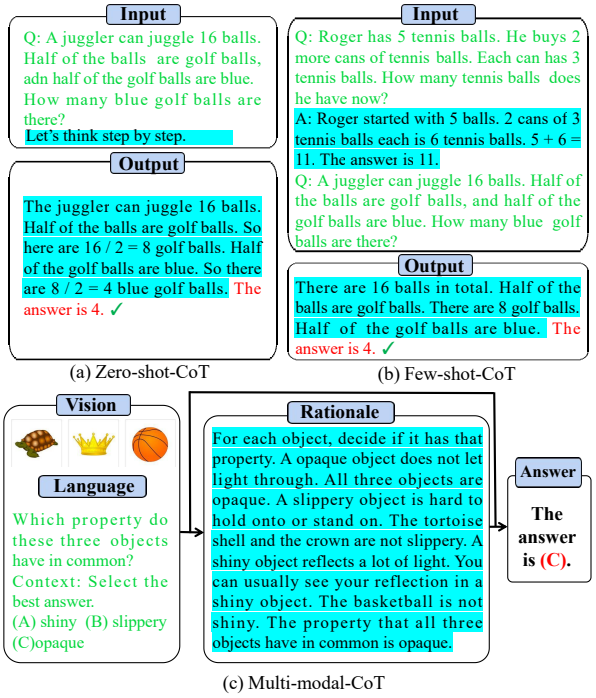

Figure 2: (a) Zero-shot-CoT (Kojima et al. 2022) (b) Few-shot-CoT (Wei et al. 2022) (c) Multi-modal-CoT (Zhang et al. 2023b)

models such as DETR (Carion et al. 2020) or CLIP (Radford et al. 2021). However, fixed image features do not align well with flexible text queries. And vision models that extract these features are not optimized for producing useful visual information that would lead to effective rationales generated by language models. For example, while DETR detects objects, its extracted features may only pay attention to the main objects in an image. Additionally, while CLIP is trained on (image, text) pairs, it only extracts shallow image information. Shallow vision features may not help language models infer correct answers because they are not closed to the reasoning. For example, as shown in Figure 1, we can not synthetize language thought if we look at the lawn in the picture and the text on the banner separately. We hypothesize that in both the stage of rationale generation and the stage of answer inference for complex problem solving, there is a need for deep understanding of visual features that capture different information in images. Therefore, effectively utilizing different modalities remains a key challenge. In this work, we propose an approach to enhance the complex reasoning ability of large language models by improving their ability to synthesize and employ language thoughts. Our approach leverages both language and vision information to achieve this goal. We propose to obtain a multi-modal latent space that deeply fuses visual features and text representations via a diffusion process. This allows our method to develop deep-level understanding, alignment and reasoning of both visual and linguistic modalities, resulting in more effective language thought generation.

Drawing inspiration from diffusion models , we employ the diffusion process to learn a text-image aligned latent space for language thought reasoning. The diffusion process entails the sequential application of multiple transformations to the latent space of image representation, where the level of noise is gradually augmented with each iteration. As a result, a series of increasingly blurred representations of the original image input is generated, ultimately leading to random noise that follows a Gaussian distribution. During each stage of the noise prediction, the model acquires a novel representation of the joint text-image distribution that captures more intricate dependencies and higher-level semantics. By repeating this procedure across several iterations, the model can acquire a deep and well-aligned latent space that encodes abundant information about both modalities. This approach is particularly useful for CoT reasoning tasks, where the goal is to reason about a long sequence of inputs and their corresponding image. By learning a deep latent space that captures high-level dependencies between text and images, it is well-suited for complex reasoning tasks.

We conducted experiments on the ScienceQA benchmark, which contains questions that require reasoning based on provided text and images. The results show that our proposed latent space learning is effective in generating useful chain of thought (CoT) and inferring correct answers. We achieved new state-of-the-art results on the ScienceQA benchmark with about only 1 billion parameters, outperforming the current SOTA baseline by 6.06% (base), 1.67% (large) respectively, and the strong ChatGPT system by 18.18% with less than 1/100th of the parameters, demonstrating the effectiveness of our approach. Our method also demonstrates strong ability in generating effective CoT, as evidenced by the ROUGE-L score of the rationales outperforming the baseline by 1.21. In addition to ScienceQA, we evaluated the effects of diffusion process for multi-modal latent space learning in multi-modal machine translation, where it also brought significant improvements. These results suggest that our proposed method is a general enhancement and can benefit the multi-modal information processing community. The code will be released at https://github.com/shimurenhlq/DPMM-COT.

Related Work

CoT Reasoning in LLMs

CoT is a widely applicable method for enhancing complex reasoning in large language models (LLMs) (Wei et al. 2022). CoT techniques assist LLMs in generating a series of logical reasoning steps, enabling them to think step by step about a question and arrive at the correct answer. CoT has significantly improved language models' performance in generating rationales and inferring accurate answers in numerous domains, including commonsense and arithmetic. In this section, we will discuss the progress made in eliciting CoT reasoning by prompting and fine-tuning language models.

For example, CoT of large language models dramatically improve the performance of large language models on arithmetic problems and symbolic reasoning. Existing CoT

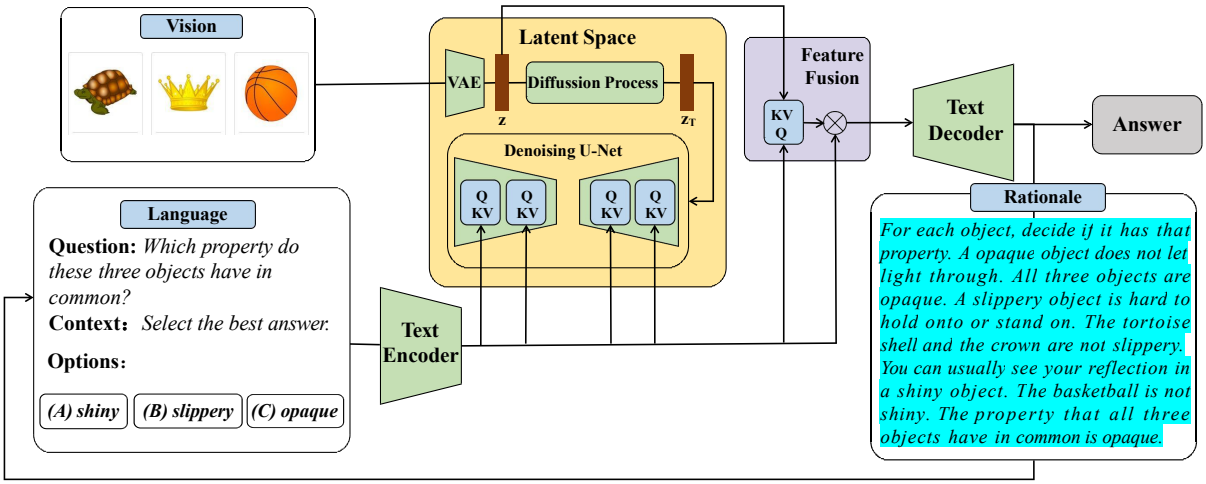

Figure 3: Overview of our multi-modal latent space learning via diffusion process for chain-of-thought reasoning in language models. Our framework consists of two stages: (i) rationale generation and (ii) answer inference.

prompting can be categorized into two major paradigms: Zero-shot-CoT (Figure 2 (a)) and Few-shot-CoT (Figure 2 (b)). Zero-shot-CoT (Kojima et al. 2022) leverages a single prompt like "Let's think step by step" to generate reasoning chains. Few-shot-CoT (Wei et al. 2022) uses reasoning demonstrations one by one. For example, prompting a PaLM 540B with eight chain-of-thought exemplars achieved state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier (Wei et al. 2022).

Multi-Modal CoT Reasoning in LLMs

To address the issue of hallucinations, which can lead to incorrect answers, and handle real-world multi-modal tasks effectively, multi-modal information can guide models to generate logical rationales. Recent studies on multi-modal Chain-of-Thought (CoT) outperform the previous state-of-the-art large language model (ChatGPT, GPT-3.5) by 16 percentage points, achieving 91.68% accuracy and even surpassing human performance on the ScienceQA benchmark (Zhang et al. 2023b).

Leveraging vision information effectively and fusing visual features with text representation in multi-modal Chain-of-Thought (CoT) poses a significant challenge. Prior work (Lu et al. 2022) has attempted to use image captions and incorporate them after text input, but this approach results in substantial information loss of images. Other studies have proposed a method that encodes texts and images using a Transformer encoder and convolutional neural network, respectively (Zhang et al. 2023a). The two sequences of representations are then fused using an attention layer for cross-modal interaction. To extract image features, Zhang et al. (Zhang et al. 2023b) employed off-the-shelf vision extraction models such as DETR (Carion et al. 2020) or CLIP (Radford et al. 2021) to obtain patch-level features and fused the information from the two modalities using an attention mechanism.

Method

In this section, we introduce our proposed Diffusion Process enhanced Multi-Modal CoT (DPMM-CoT) method. We follow the Multi-Modal CoT (MM-CoT) approach proposed by Zhang et al. (Zhang et al. 2023b) as our baseline. The overview of our full model is illustrated in Figure 3.

Multi-modal CoT

Task Definition In multi-modal reasoning, a language input including a question X_Q, its language context X_L, the options X_O, and its corresponding image X_V are usually given as inputs to the model. The model is required to answer a question X_Q according to the options X_O to obtain the answer Y_A . In other words, the model is trained to maximize the likelihood between the predicted answer $\hat{Y_A}$ and the true answer Y_A distribution. MM-CoT breaks down this problem into two steps through the introduction of a rationale Y_R: rationale generation and answer inference. In the rationale generation step, the model is required to predict a rationale $\hat{Y_R}$ that can infer the answer, which maximizes the likelihood between predictions and the standard rationale R distribution. Then, in the second stage, based on the rationale $\hat{Y_R}$, along with the language input including the question X_Q, its language context X_L, the options X_O, and the corresponding image X_V, the model predicts the final answer Y_A.

Text Encoder For the multi-modal CoT tasks, the text input differs between the two stages. In the stage of rationale generation, the text input includes language context X_L, a question X_Q, and multiple options X_O. In the stage of answer inference, the text input comprises language context X_L, a question X_Q, multiple options X_O, and rationale $\hat{Y_R}$ generated from the first stage. We adopt the Transformer model for text encoding, which is initialized by the pre-trained model UnifiedQA (Khashabi et al. 2020). We obtain

the text representation Z_L as follows:

$$Z_L^R = \text{ENCODER}_{text}([X_L; X_Q; X_O]), \quad (1)$$

$$Z_L^A = \text{ENCODER}_{text}([X_L; X_Q; X_O, \hat{Y_R}]), \quad (2)$$

where $Z_L = Z_L^R$ in the stage of rationale generation and $Z_L = Z_L^A$ in the stage of answer inference.

Image Feature Extraction and Feature Fusion In multi-modal CoT, the image encoder plays a crucial role in the CoT process as it helps to provide additional context and information to the model. By incorporating visual features extracted from input images, the model gains a better understanding of the overall context. Specifically, the image feature Z_V is first extracted by an image encoder:

$$Z_V = \text{ENCODER}_{img}(X_V). \quad (3)$$

Based on the acquired image features, in order to integrate the image and text encoding features, a linear layer is first used to map the image features. This is primarily for two purposes: to unify the dimensions of image and text features, and to project the image features onto the same feature space that can be fused with text features.

$$Z_V^T = W_h * Z_V, \quad (4)$$

where W_h is the learnable weight matrix.

As the image features and text features have different temporal lengths, we use an attention mechanism to project the image features onto the length of the text features based on the correlation between the image and text features. Specifically, we use Z_L as the attention query , with Z_V^T as attention keys and values. The resultant projected image features are as follows:

$$Z_V^{attn} = Softmax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}})\mathcal{V} \quad (5)$$

where d_k is the dimension of Z_L, $\mathcal{Q}$ is the Z_L, $\mathcal{K}$ and $\mathcal{V}$ are Z_V^T.

As the roles of image features and text features in generating rationales and answers are not static or fixed, we choose to use a gate mechanism to fuse vision features and language representation, i.e., let the model decide how to use the image and text features. The gated fusion mechanism ((Zhang et al. 2020) (Wu et al. 2021) (Li et al. 2022)) involves two steps: obtaining a score vector between 0 and 1 to determine the importance of each feature (Equation 6), and using the scores to fuse the text and attention features (Equation 7).

$$\alpha = Sigmoid(W_l Z_L + W_v Z_V^{attn}), \quad (6)$$

$$Z_{fuse} = (1 - \alpha) * Z_L + \alpha * Z_V^{attn} \quad (7)$$

where W_l and W_v are learnable parameters for gate projection, $Z_{fuse} = Z_{fuse}^R$ in the stage of rationale generation and $Z_{fuse} = Z_{fuse}^A$ in the stage of answer inference.

Text Decoder In multi-modal CoT, the text decoder is responsible for generating rationales or inferring the final answers, taking into account the representation output Z_{fuse} of the encoder and the previously decoded token to predict the next one. For example, in the stage of rationale generation, the decoder predicts the rationale $Y = (y_1, \ldots, y_N)$ token by token, according to the last decoding state and source context. The rationale probability can be formulated as follows:

$$s_i = \text{SELFATTN}(Y_{<i}), \quad (8)$$

$$P(y_i|Y_{<i}, Z_{fuse}; \theta) = Softmax(FFN(s_i + \text{CROSSATTN}(s_i, Z_{fuse}))), \quad (9)$$

where θ is the model parameters, y_i is the i-th token in Y with N tokens, s_i denotes the decoding state at the i-th timestep.

Therefore, the sequence generation loss $\mathcal{L}_{SEQ}$ for model optimization can be written as:

$$\mathcal{L}_{SEQ} = \sum_{i=1}^{N} -\log P(y_i|Y_{<i}, Z_{fuse}; \theta) \quad (10)$$

where $Y = Y_R$ in the stage of rationale generation, and $Y = Y_A$ in the stage of answer inference.

Multi-modal Latent Space Learning

In the current multi-modal CoT work, such as MM-CoT (Zhang et al. 2023b), image feature extraction is performed using off-the-shelf image encoders trained on models such as DETR (Carion et al. 2020) and CLIP (Radford et al. 2021). However, this method has two major drawbacks. Firstly, due to the limitations of pre-training objectives, the extracted image features are usually shallow and generic information that is not specifically optimized for reasoning, thus lacking deep semantic information which is required in reasoning. Secondly, the image features used for inference are highly dependent on language input, meaning that different image features are required for different language descriptions. Therefore, this work proposes a method of multi-modal latent space learning, which learns flexible image features that are aligned with text inputs in the latent space and optimized for the inference process, thus possessing the deep semantics required for reasoning.

As Richard Feynman once said, "*What I cannot create, I do not understand*." Therefore, we argue that excellent creativity must contain excellent understanding. Drawing inspiration from the outstanding generative performance of diffusion models, we apply the idea of a stable diffusion model to obtain a multi-modal latent space. Specifically, we use the concept of a diffusion process to obtain better image features with deep semantics that align with text representation. Firstly, we employ the Variational AutoEncoder (VAE) (Kingma and Welling 2014) as the image encoder to obtain the latent vector of the image. Then, we add random noise to the latent vector, which follows a Gaussian distribution with time steps. Next, the latent vector is inputted into the UNet neural network (Ronneberger, Fischer, and Brox 2015). We fuse text representation and image features at a deep level by mapping text information into the intermediate layer of UNet through a cross-attention layer. By optimizing the diffusion process, which compares the predicted noise with true noise, the model can obtain better image features with deep

semantics that align with text inputs. This is because the diffusion process enables the model to learn features that are not only optimized for reasoning but also possess a high degree of stability and robustness to noise and other disturbances. In this way, the model can obtain the latent space of an image with deep semantics from the perspective of diffusion.

Stable diffusion consists of two main parts: (1) the forward (or diffusion) process and the reverse process. In the diffusion process, random noise following a Gaussian distribution is added to the image features of latent space. This process is entirely run on the latent space and is composed of a VAE neural network and a scheduling algorithm. (2) The reverse process generates an image using an image decoder based on the latent features and text representations. Existing studies (Kwon, Jeong, and Uh 2022) have shown that the latent space already contains aligned semantic information, so we suppose that it can be utilized to fuse linguistic modality and visual modality for reasoning.

In our DPMM-CoT, we first encode the image into the latent space Z_V^0 using VAE. Especially, during inference, the image is encoded as a latent vector through the VAE and then directly fused with the text representation vector to generate a rational or answer. Then we add random noise that follows a Gaussian distribution to the latent space of the image.

$$Z_V^0 = \eta VAE(X_V), \tag{11}$$

$$q(Z_V^t|Z_V^{t-1}) = \mathcal{N}(Z_V^t; \sqrt{1-\beta_t}Z_V^{t-1}, \beta_t\mathbf{I}) \tag{12}$$

which indicates the diffusion process that adding noise that follows a Gaussian distribution, where β_t is the variance schedule, $\sqrt{1-\beta_t}Z_V^{t-1}$ is the mean, $\mathbf{I}$ is identity matrix, and $\eta = 0.18215$ is the scale factor.

The diffusion process of the diffusion model can be expressed as a Markov chain from $t = 0$ to $t = T$:

$$q(Z_V^{0:T}) = q(Z_V^0)\prod_{t=1}^{T} q(Z_V^t|Z_V^{t-1}). \tag{13}$$

When $T \to \infty$, the final result will become a noisy image, similar to sampling from an isotropic Gaussian distribution. However, we also use a closed-form formula to directly sample noisy images at a specific time step t, instead of designing an algorithm to iteratively add noise to the image following the practice of (Rombach et al. 2022).

$$Z_V^t = \sqrt{\overline{\alpha}_t}Z_V^0 + \sqrt{1-\overline{\alpha}_t}\epsilon \tag{14}$$

where α_t = 1 - β_t, $\overline{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. ϵ is an i.i.d. (independent identically distributed) standard normal random variable. It is important to distinguish them using different symbols and subscripts because they are independent and their values may differ after sampling.

The standard diffusion process involves predicting the noise using UNet (Ronneberger, Fischer, and Brox 2015). By utilizing a cross-attention layer to map text information into the intermediate layer of UNet, we can merge text representation with image features. This integration of information from both modalities leads to a more comprehensive understanding of the underlying structure in the data. Text features provide valuable additional semantic information that may not be immediately evident from the visual content alone. Incorporating these features into the model allows us to better comprehend the context and meaning behind the visual elements. Meanwhile, image features offer rich visual information about the objects and scenes depicted in the image. These features enable the model to identify patterns and relationships between different parts of the image. The fusion of both image and text features through diffusion process enables the model to leverage the strengths of both modalities, leading to improved latent space learning.

Specifically, we predict the noise $\epsilon_\theta(Z_V^t, t, Z_L)$ by UNet with the attention mechanism between visual feature Z_V^t and text representation Z_L as follows:

$$\epsilon_\theta(Z_V^t, t, Z_L) = \text{UNET}(FFN(Softmax(\frac{QK^T}{\sqrt{d}})V + Q)) \tag{15}$$

where $Q = W_Q^{(i)} \cdot Z_V^t$, $K = W_K^{(i)} \cdot Z_L$, $V = W_V^{(i)} \cdot Z_L$. $Z_V^t \in \mathbb{R}^{N\times d^i}$ is an intermediate representation of UNet. Therefore, the latent diffusion process loss implemented with Maximum Square Error (MSE) can be written as:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0.1),Z_L,t}[||\epsilon - \epsilon_\theta(Z_V^t, t, Z_L)||_2^2] \tag{16}$$

where $\mathcal{L}_{LDM}$ is the loss of latent diffusion model, θ is the parameters of the model, ϵ is the random noise (an independent identically distributed standard normal random variable). So the total loss of the model is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{SEQ} + \mathcal{L}_{LDM}. \tag{17}$$

During the training process, all model parameters, except for the UNet parameters in the reverse process, are updated.

Experiments

Setup

Datasets To assess CoT on LLMs, we followed the approach of MM-CoT (Zhang et al. 2023b) and used the Science Question Answering (ScienceQA) (Lu et al. 2022) dataset. In addition, we also conducted experiments on the Multi30K multi-modal translation dataset (Elliott et al. 2016) and followed the work of IKD-MMT (Peng, Zeng, and Zhao 2022).

Settings In our experiment on CoT in LLMs, we employed a two-stage framework consisting of two procedures: rationale generation and answer inference. Both stages shared the same model architecture, namely the T5 encoder-decoder architecture (Raffel et al. 2020).

For our experiment on multi-modal machine translation (MMT), we employed the mT5 encoder-decoder architecture, which was initialized using the pre-trained mT5-large (Xue et al. 2021) checkpoint, which had been pre-trained on a multilingual corpus.

Main Analysis

Table 1 presents the main results of our study, which compares the performance of various Visual Question Answering (VQA) models. We evaluated our DPMM-CoT

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
ViLT (Kim, Son, and Kim 2021)	113M	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90	61.14
Patch-TRM (Lu et al. 2021)	90M	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50	61.42
VisualBERT (Li et al. 2019)	111M	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA$_{\text{Base}}$ (Khashabi et al. 2020)	223M	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00	70.12
UnifiedQA$_{\text{Base}}$ w/ CoT (Lu et al. 2022)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
ChatGPT (GPT-3.5)	175B	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
ChatGPT (GPT-3.5) w/ CoT (Lu et al. 2022)	175B	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
MM-CoT$_{\text{Base}}$(CLIP)	223M+151M	87.97	80.88	87.36	88.32	84.78	88.15	86.34	86.29	86.32
MM-CoT$_{\text{Base}}$(DETR) (Zhang et al. 2023b)	223M+60M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
DPMM-CoT$_{\text{Base}}$	223M+83M	**92.72**	**87.85**	**89.91**	**92.72**	**90.48**	**91.29**	**91.45**	**90.11**	**90.97**
MM-CoT$_{\text{Large}}$(DETR) (Zhang et al. 2023b)	738M+60M	**95.91**	82.00	90.82	95.26	88.80	**92.89**	92.44	90.31	91.68
DPMM-CoT$_{\text{Large}}$	738M+83M	95.52	**90.33**	**91.36**	**95.50**	**93.26**	92.68	**93.28**	**93.47**	**93.35**

Table 1: Main results on ScienceQA test set (%). 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. Results except ours are taken from (Lu et al. 2022) and (Zhang et al. 2023b).

Method	(i) QCM→ R	(ii) QCMR→ A
MM-CoT$_{\text{Base}}$	96.97	84.91
DPMM-CoT$_{\text{Base}}$	98.18	90.97

Table 2: Performance of two-stage.

Model	EN-DE			EN-FR	
	Test16	Test17	MSCOCO	Test16	Test17
IKD-MMT	41.28	33.83	30.17	62.53	54.84
mT5	38.56	33.01	28.10	61.71	53.84
DPMM-MT	41.63	**36.18**	**30.75**	**66.91**	**57.80**

Table 3: BLEU score of EN-DE and EN-FR tasks.

model against MM-CoT, a baseline, and found that DPMM-CoT$_{\text{Base}}$ outperforms MM-CoT$_{\text{Base}}$(DETR) by 6.06% and DPMM-CoT$_{\text{Large}}$ outperforms MM-CoT$_{\text{Large}}$(DETR) by 1.67%. Notably, when questions involve visual context (IMG column), DPMM-CoT$_{\text{Base}}$ and DPMM-CoT$_{\text{Large}}$ outperform MM-CoT$_{Base}$(DETR) and MM-CoT$_{\text{Large}}$(DETR) by 7.58% and 4.46%, respectively.

Compared to other VQA baselines, DPMM-CoT$_{\text{Large}}$ outperforms VisualBERT (Li et al. 2019) by 31.48%, demonstrating that autoregressive language pre-training and larger language models are effective for problem solving. And DPMM-CoT$_{\text{Large}}$ surpasses the UnifiedQA model with CoT (Lu et al. 2022) by 19.24%. This suggests that only leveraging captions of images as visual context causes severe information loss and hallucination in CoT.

Additionally, we found that DPMM-CoT$_{\text{Large}}$ outperforms the strong LLM – ChatGPT by 18.18%, demonstrating that language models under 1B parameters can perform better than general LLMs when fine-tuned with appropriate network designs and information. Moreover, our DPMM-CoT$_{\text{Base}}$ and DPMM-CoT$_{\text{Large}}$ both outperform human performance, indicating the effectiveness of our model. These results suggest that multi-modal latent space learning is significant for understanding flexible and deep visual information. In Table 2, the ROUGE-L results of rationals generated by DPMM-CoT$_{\text{Base}}$ and MM-CoT$_{\text{Base}}$, as well as the accuracy of answers inferred, are shown.

To verify that the improvement of DPMM-CoT$_{\text{Large}}$ originates from multi-modal latent space learning via the diffusion process rather than an increase in the number of parameters, we utilized fixed visual features extracted by clip-vit-base-patch32 (Zhang et al. 2023b), which has 151M parameters. The result shows that while an increase in the number of parameters may contribute to improved performance on multi-modal QA tasks, it is still far from our DPMM-CoT model. This suggests that our improvements are due to a deeper understanding of visual information gained through multi-modal latent space learning.

Further Analysis

Generalization to More Multi-modal Tasks To demonstrate the generality of our method across different multi-modal tasks, we conducted experiments on Multimodal Machine Translation (MMT). The main results are presented in Table 3. We trained our Diffusion Process Enhanced Multi-Modal Machine Translation (DPMM-MT) model on the Multi30K dataset, which includes English-to-French and English-to-German translations. We then evaluated DPMM-MT on three test sets: test2016-flickr (Test16), test2017-flickr (Test17), and test2017-mscoco (MSCOCO). Firstly, compared to the mT5 baseline that does not use image features, we achieved significant improvements in En-De of 3.07, 3.17 and 2.65, and in En-Fr of 5.20 and 3.96, respectively. This indicates the crucial role of image features in multimodal machine translation. We achieved a new state-of-the-art (SOTA) result on Test17 and MSCOCO with English-to-German translation and Test16 and Test17 with English-to-French translation. Specifically, DPMM-MT outperformed the previous SOTA by 2.35 (33.83 →

Model	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Zero Tensor	92.72	87.85	89.91	92.72	90.48	91.29	91.45	90.11	90.97
Blank Image	92.54	82.56	89.91	92.86	88.35	90.94	90.68	88.13	89.77

Table 4: Results of different way of solving problems without images.

Model	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Our model	92.72	87.85	89.91	92.72	90.48	91.29	91.45	90.11	90.97
w/o Stable Diffusion Pre-training	88.63	80.43	85.45	89.93	84.88	85.92	87.15	84.18	86.09
w/o UNet	91.92	82.56	89.91	92.18	88.35	90.52	89.98	88.46	89.44
w/ Frozen VAE	91.07	82.00	90.36	91.64	87.36	90.73	89.35	88.33	88.99

Table 5: Ablation results of our method.

36.18) and 0.58 (30.17 → 30.75) on Test17 and MSCOCO with English-to-German translation, respectively. DPMM-MT outperformed the previous SOTA by 4.38 (62.53 → 66.91) and 2.96 (54.84 → 57.80) on Test16 and Test17 with English-to-French translation, respectively. For Test16 with English-to-German translation, we also achieved comparable results to the previous SOTA - Gated Fusion (Wu et al. 2021). These improvements across multiple datasets suggest that utilizing our proposed multi-modal latent space learning to extract deep image semantics is useful for enhancing the performance of multi-modal machine translation.

Problems without Images Since not all questions in the ScienceQA task (or other real-life tasks) include images, our method needed to be adaptable to image-less questions. For this purpose, we explored two approaches: using blank images or null tensors as input for these questions. We analyzed the results using models DPMM-CoT$_{\text{Base}}$, and the experiment outcomes are presented in Table 4. Our findings show that using zero tensors resulted in a 1.20% higher accuracy than using blank images. This may be attributed that blank images may introduce misleading information during the diffusion process.

Ablation Study

To illustrate the effect of each component in the Diffusion Process on multi-modal latent space learning, we conducted an ablation study. As shown in Table 5, we tested whether pre-trained stable diffusion module is useful for multi-modal latent space learning. We randomly initialized the parameters of UNet and VAE, and evaluated the result without Stable Diffusion Pre-training. The results show that diffusion models including VAE and UNet initialized from pre-trained model are indeed useful for DPMM-CoT. The accuracy declined by 4.88% (90.97% → 86.09%), demonstrating the importance of good initialization for producing effective multi-modal latent space. Furthermore, we found that diffusion components initialized by random parameters actually outperform the baseline MM-CoT(DETR). This highlights the ability of the diffusion process to deeply understand image information after being trained on the ScienceQA dataset, producing effective image features aligned with text representation.

To further demonstrate the importance of diffusion process in multi-modal latent space learning, we trained the model without UNet. The images were only encoded by VAE to produce latents. The result in Table 5 shows that accuracy declined by 1.53% (90.97% → 89.44%), indicating the significance of diffusion process to produce multi-modal latent space. These results testify that diffusion process is a key part of multi-modal latent space learning, and visual feature extraction by encoder alone is insufficient. By adding noise and predicting noise with UNet guided by text representation, the multi-modal latent space learning gains a deep understanding of image with language thoughts.

The quality of the vision latent vector that VAE produces has a significant impact on the effectiveness of the CoT. To prove this, we tried not updating the parameters of VAE during CoT training but instead used pre-trained parameters from Stable-Diffusion-v1-4. The results (90.97% → 88.99%) show that VAE trained with CoT is helpful in producing better latent vectors for use in reasoning. This also demonstrates that for reasoning tasks, it's not enough to perform only self-supervised pre-training.

Conclusion

In this work, we focuses on improving the production of multi-modal latent spaces that can effectively understand both linguistic and visual information at a deeper level. To achieve this, we introduce DPMM-CoT, a multi-modal latent space learning approach via diffusion process for CoT reasoning in language models. Our experimental results demonstrate that our method performs exceptionally well on multi-modal tasks. Notably, DPMM-CoT$_{\text{Base}}$ outperforms MM-CoT$_{\text{Base}}$ by 6.06%, while DPMM-CoT$_{\text{Large}}$ outperforms MM-CoT$_{\text{Large}}$ by 1.67%. We also conducted additional experiments on multi-modal machine translation, which verified the effectiveness of our proposed multi-modal latent space learning method on a wider range of multi-modal tasks. Moreover, our concrete analysis shows that our method enables language models to attain deeper, more flexible, and aligned features for language thought, thereby enhancing their reasoning abilities. In the future, we plan to evaluate our method on more multi-modal tasks.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, 213–229. Springer.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 70–74. Berlin, Germany: Association for Computational Linguistics.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1896–1907. Online: Association for Computational Linguistics.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *CoRR*, abs/2205.11916.

Kwon, M.; Jeong, J.; and Uh, Y. 2022. Diffusion Models already have a Semantic Latent Space. *CoRR*, abs/2210.10960.

Li, B.; Lv, C.; Zhou, Z.; Zhou, T.; Xiao, T.; Ma, A.; and Zhu, J. 2022. On Vision Features in Multimodal Machine Translation. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 6327–6337. Association for Computational Linguistics.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.; Zhu, S.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *CoRR*, abs/2209.09513.

Lu, P.; Qiu, L.; Chen, J.; Xia, T.; Zhao, Y.; Zhang, W.; Yu, Z.; Liang, X.; and Zhu, S.-C. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Peng, R.; Zeng, Y.; and Zhao, J. 2022. Distill The Image to Nowhere: Inversion Knowledge Distillation for Multimodal Machine Translation. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 2379–2390. Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10674–10685. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N.; Hornegger, J.; III, W. M. W.; and Frangi, A. F., eds., *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E. H.; Le, Q.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *CoRR*, abs/2201.11903.

Wu, Z.; Kong, L.; Bi, W.; Li, X.; and Kao, B. 2021. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. *arXiv preprint arXiv:2105.14462*.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 483–498. Association for Computational Linguistics.

Zhang, Z.; Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Li, Z.; and Zhao, H. 2020. Neural Machine Translation with Universal Visual Representation. In *International Conference on Learning Representations*.

Zhang, Z.; Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Li, Z.; and Zhao, H. 2023a. Universal Multimodal Representation for Language Understanding. *CoRR*, abs/2301.03344.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023b. Multimodal Chain-of-Thought Reasoning in Language Models. *CoRR*, abs/2302.00923.