Robust Reasoning with Contextualized Visual Representation Learning

Anonymous ACL submission

Abstract

Visual question answering (VQA) requires vision-language models (VLMs) to reason over images and respond to questions that ask about diverse details and inferences of these images. Typically, VLMs use pre-trained vision encoders to map visual inputs to feature representations, and fuse these representations with large language models (LLMs), which generate responses to questions. However, these query-agnostic visual representations only reflect a static set of features of the visual input, which hinders VLMs from robustly responding to queries about out-of-distribution (OOD) features. To address this challenge, we propose to fuse the query as additional context into early-stage vision encoding, enabling models to learn context-aware visual representations that can flexibly adapt to different queries. Our contextualized vision transformer, C-ViT, learns the early fusion of vision and context via a fine-grained curriculum learning scheme, based on a novel Contextual Vision-Inference Alignment (CVIA) dataset. We apply C-ViT to two VLM architectures, and results on both architectures demonstrate that C-ViT effectively improves reasoning robustness of VLMs, particularly when generalizing to OOD VQA data.

1 Introduction

Visual question answering (VQA; Singh et al., 2019) requires vision-language models (VLMs) to reason over visual inputs (images) with text-based queries, and generate appropriate textual responses. To answer visual questions, VLMs typically use pre-trained vision encoders to convert visual inputs to feature representations. These features are then concatenated with query embeddings, to enable reasoning across modalities via large language models (LLMs) (Liu et al., 2023a; Wang et al., 2024; Ghosh et al., 2024). However, the pre-trained vision encoder does not receive the queries as contexts when constructing the visual feature representations. As
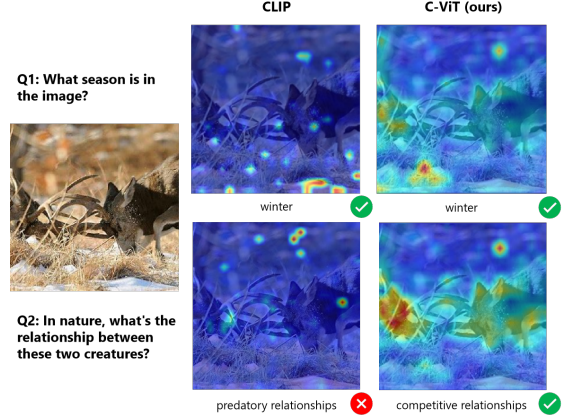

Figure 1: Illustration of how our contextualized visual representation dynamically adapts to different contexts (queries). Heatmap areas with warmer colors indicate image regions that the model puts more attention to when responding to each query. LLaVA-1.5 with our C-ViT contextual vision encoder shows more flexible attention to image regions that are relevant to each query.

a result, current VLMs that rely on such context-agnostic representations may fail to generalize to out-of-distribution (OOD) queries (Mayilvahanan et al., 2024b; Liu et al., 2024b; Abbasi et al., 2024), which may require visual feature representations that are more fine-grained than the ones yielded by a query-agnostic visual encoder.

For example, as shown in Figure 1, to answer question Q1, which asks about the season presented in the image, the model must focus on the snow at the bottom part of the image. In contrast, for answering question Q2 that concerns the relationship between the creatures, the model needs to pay attention to the center of the image instead, where two creatures are engaged in a fight. VLMs with context-agnostic vision encoders such as CLIP (Radford et al., 2021) only capture visual features that are relevant to Q1, failing to extract features that are useful for answering Q2. This loss of visual information during early-stage encoding reduces

the reasoning robustness of VLMs when encountering various real-world input queries.

In this work, we propose a **C**ontextualized **Vi**sion **T**ransformer (**C-ViT**) to overcome the limitations of context-agnostic visual representations, by incorporating the input query (as additional context) into the early stage of vision encoding. C-ViT's early fusion of vision and query enables the learning of context-aware visual representations, which dynamically produce visual features that are more relevant to varied input queries, enhancing VLMs' reasoning robustness, as seen in Figure 1.

To facilitate the learning of contextualized visual representations, we introduce a **C**ontextual **V**ision-**I**nference **A**lignment (**CVIA**) dataset along with its automatic generation pipeline. CVIA contains 10M image–context–inference triples, which form contrastive learning samples on three difficulty levels (basic, easy and hard), according to the difficulty of discriminating the positive versus negatives and the number of negatives included per sample. Based on CVIA, we propose a three-stage curriculum learning pipeline to train C-ViT, which progressively enhances the alignment of C-ViT's context-aware vision encoding with the output inference corresponding to the context. Specifically, in the first two stages, to learn the alignment, C-ViT is trained on our contrastive learning data CVIA gradually from basic, easy to hard levels. While in the final stage, we finetune the entire VLM (*i.e.*, LLM plugged with C-ViT) end-to-end on downstream VQA tasks, to further adapt C-ViT's encoding into the LLM inference process.

We evaluate C-ViT by integrating it into two leading VLM architectures: the encoder-decoder framework (Cho et al., 2021) and the decoder-only framework (Liu et al., 2024a). The models are benchmarked on a diverse set of vision-language tasks. Across all benchmarks, models equipped with C-ViT consistently outperform their baseline vision encoders, particularly in zero-shot settings. Our results highlight C-ViT's robustness to domain shifts and its strong generalization to OOD contexts. This performance is attributed to C-ViT's dynamic focus on contextually relevant regions of an image, as verified by our analysis in Section 6.3.

2 Background: Early-fusion Vision Encoder

Traditional late-fusion architectures (Bai et al., 2025; Chen et al., 2024b) combine visual features and text embeddings at the stage just before the final LLM inference generation. In doing so, they produce a context-agnostic image representation that related only to image itself, constraining its ability to extract task-specific details and, consequently, hindering performance on complex vision-language tasks (Li et al., 2024a).

In contrast, early-fusion architectures (Team, 2024) introduce contextual information at the early stage of image feature processing, leading to dynamic vision representations that adapt to the accompanying context. By conditioning visual features in context, early-fusion architectures exhibit greater robustness and generalization ability, as verified in Section 5.

QA-ViT (Ganz et al., 2024) accomplishes early-fusion by augmenting a frozen CLIP vision encoder to form a conditional encoder, $V(I) \rightarrow V(I|Q)$.

The input question Q is encoded to produce textual features F_Q via the LLM's text encoder. At each self-attention block of the vision encoder, these question features are concatenated with the visual sequence F_V, enabling a cross-modal attention over the combined sequence. The output corresponding to the visual sequence F'_{VQ} is able to capture context-aware vision features:

$$F'_{VQ} = Attention(concat(F_V, F_Q))_{[0:M]} \quad (1)$$

where M is the length of vision sequence. Finally, the fused vision feature passes through both the original projection P and a learnable gated projection P_g, maintaining the layer's outputs with minimal deviation at initialization while enabling a residual learnable stream of information:

$$F_{VQ} = P(F'_{VQ}) + P_g(F'_{VQ}) \cdot tanh(\beta) \quad (2)$$

In this work, we follow QA-ViT to build a early-fusion contextual vision encoder C-ViT, and propose a method to seek stronger learning of contextualized visual representations using C-ViT.

3 Contextualized Visual Representation Learning

In this section, we propose a novel method for contextualized visual representation learning. First, we introduce the Contextualized Vision Transformer (C-ViT), a vision encoder that integrates contextual information from the early-stage of vision encoding. Next, we describe our Contextual Vision–Inference Alignment dataset (CVIA), designed to align C-ViT's outputs with visual inferences. Finally, we detail a curriculum learning
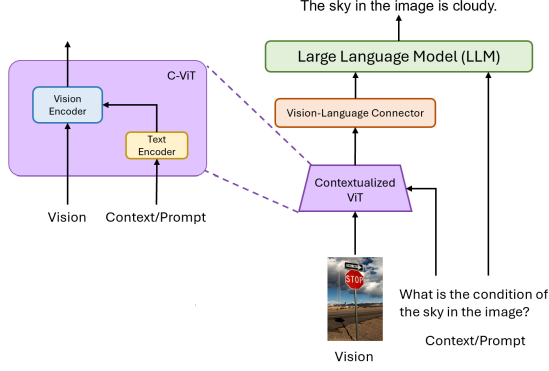
Figure 2: An overview of our C-ViT architecture. C-ViT consists a vision and a text encoder to process both image and text inputs, jointly generating contextualized visual representations. It enables cross-modal interaction by injecting context features directly into each self-attention layer.

pipeline that progressively increases sample difficulty throughout training.

3.1 Contextualized Vision Transformer

Building on QA-ViT architecture described in Section 2, we propose the Contextualized Vision Transformer (C-ViT), a contextual vision encoder designed to early fuse the input context (query) with the visual features, as illustrated in Figure 2. Unlike late-fusion approaches, which combine modalities only at LLM input stage, C-ViT injects context features into the self-attention layers of the vision encoder. This allows rich cross-modal interactions, leading to a context-aware representation that provides more effective feature extraction.

C-ViT consists of two main components: a vision encoder and a text encoder. The vision encoder is responsible for generating visual representations and the text encoder provides the contextual representations needed. The exact implementation of C-ViT can vary depending on the VLM architecture. Specifically, the text encoder can inherit the LLM's encoder in an encoder-decoder architecture, while in a decoder-only setup, it's initialized by a pretrained CLIP text encoder.

Thanks to this flexibility, C-ViT functions as a plug-and-play module, supporting both encoder-decoder and decoder-only architectures.

3.2 Contextual Vision-Inference Alignment Dataset

Recent findings (Li et al., 2024b) suggest that vision and language models learn similar representations of the world, differing only in spatial distribu-

tion. The cross-modality alignment mainly relies on the vision-language connector in a VLM. Since C-ViT modifies the original CLIP vision encoder architecture by introducing new input and parameters, the original output distribution is disrupted. As a result, we must retrain C-ViT to learn the correct representations. To support this, we propose a Contextual Vision-Inference Alignment (CVIA) dataset, which contains three types of samples and supports curriculum contrastive learning.

Figure 3: An example from our CVIA. The dataset is constructed for curriculum contrastive learning, containing positive samples paired with negative samples across three difficulty levels: basic, easy, and hard. Each level includes ten negative samples per positive instance.

The CVIA dataset categorizes samples into three difficulty levels: basic, easy, and hard. Basic samples are obtained via random negative sampling from the Cambrian dataset (Tong et al., 2024), while easy and hard samples are generated by our proposed data-augmentation strategies: negative image augmentation and negative context augmentation, respectively. An example is illustrated in Figure 3. We further denote our dataset as CVIA$_{easy}$ or CVIA$_{hard}$ when the majority of its samples come from the easy or hard level, respectively. Our design allows precise control over training difficulty by tuning the proportions of basic, easy, and hard samples. Detailed implementations of these strategies are provided in the following sections.

3.2.1 Negative Image Augmentation

For each sample, our objective is to identify a different image that yields a different visual inference when paired with the same context. To achieve this, we search for negative images within the Cambrian dataset by computing CLIP embeddings and selecting those that have high cosine similarity scores to the positive image. We then apply a two-round verification with a VLM to ensure that the positive
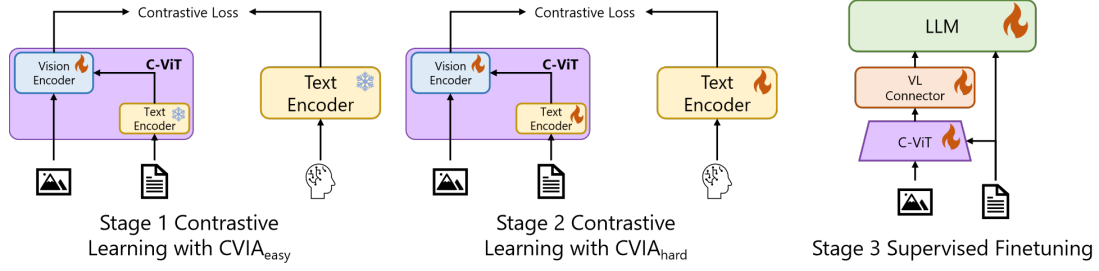
Figure 4: An overview of the three-stage curriculum learning pipeline. Stage 1 and Stage 2 use curriculum contrastive learning to finetune C-ViT using CVIA dataset under two difficulty settings. Stage 3 use supervised finetuning to finetune the entire VLM architecture.

inference is not a plausible answer for the negative images, and vice versa. Finally, we employ a LLM to rewrite each negative inference to eliminate any length-based shortcuts.

We label the data produced by this strategy as "easy" negatives, since each positive sample can only be paired with exactly one negative in one contrastive batch, because there is no guarantee that different negatives will be mutually irrelevant. We include our detailed prompts for sampling negative images in Appendix B.1 and B.3.

3.2.2 Negative Context Augmentation

Negative context augmentation aims to pair each image with a diverse set of contexts. Although the Cambrian dataset includes several contexts per image, we enrich this pool by following prior work (Zhu et al., 2024) and implementing an automated questioning pipeline to generate additional negative contexts and their corresponding inferences.

Our pipeline utilizes two separate VLMs: a questioner, which formulates questions about the image, and an answerer, which provides the visual inferences. The pipeline proceeds in three levels (coarse-grained, fine-grained, and reasoning-required) of question generation. The questioner moves forward to the next level once no new questions can be generated on the current topic. Additionally, we still deploy a LLM to complete the rewriting process to prevent shortcuts.

We label the resulting samples as "hard" negatives, since each image can be paired with multiple, mutually irrelevant contexts, forcing the model to learn stronger discriminative capabilities when all negatives are included in the same contrastive batch. The detailed prompts for negative context sampling can be found in Appendix B.2.

3.3 Curriculum Learning Pipeline

Curriculum learning (Wang et al., 2021) introduces training samples in increasing order of difficulty, enabling models to first master simpler examples before tackling more complex ones. This gradual progression has been shown to improve convergence and final performance, particularly in transfer learning scenarios where smoother adaptation to new tasks is critical.

Inspired by this approach, we designed a three-stage curriculum learning pipeline (Figure 4) with two distinct training objectives: aligning visual features with contextual understanding and enhancing performance on downstream tasks. The first two stages focus on curriculum contrastive learning, training the model to align visual representations with their corresponding contextual inferences. In the final stage, we apply supervised finetuning to the entire VLM, allowing us to evaluate its effectiveness on downstream tasks.

We believe that this staged approach fosters a strong generalization capability, enabling C-ViT to function as a strong and flexible vision encoder when plugged into a VLM. The detailed strategy for each stage is described as follows.

3.3.1 Contrastive Learning Stages

Combined with C-ViT, we introduce a novel contrastive learning objective that requires three inputs: image, context, and visual inference. The goal is to align the visual representation, conditioned on both the image and context, with the representation of its corresponding visual inference. We employ a two-stage training strategy that progressively increases data difficulty while gradually unfreezing model parameters, with a smooth transition from the pretrained CLIP encoder to our new objective.

Stage One In the first stage, we train on $CVIA_{easy}$. At this stage, the text encoder is kept

4

frozen and only the newly added vision-encoder parameters are updated. This configuration enables the model to begin embedding contextual information into its visual representations while preserving CLIP's foundational encoding capabilities. By starting with simpler examples and limiting the scope of parameter updates, we reduce the risk of overfitting or shortcut learning, and ensure a stable shift from CLIP's original contrastive objectives to our context-aware targets.

Stage Two   In the second stage, we train on $CVIA_{hard}$. Here, we unfreeze the text encoder so it can fully participate in the joint contrastive objective. This stage mainly focuses on distinguishing representations of the same image under different contexts, reinforcing the model's ability to infer how context shapes visual interpretation. By gradually increasing both data difficulty and model flexibility, our two-stage contrastive curriculum enables C-ViT to learn a rich, context-sensitive visual embedding, laying a robust foundation for subsequent downstream fine-tuning.

3.3.2   Supervised Finetuning Stage

In the third stage, we plug C-ViT into the VLM framework and perform supervised finetuning on downstream tasks. Unlike approaches such as LLaVA, which freeze the vision encoder at this point, we choose to keep C-ViT's newly introduced parameters unfrozen. This decision stems from our belief that contrastive learning with our limited data size may not be sufficient for the model to fully master how to utilize contextual information. Continuing to finetune the vision encoder allows it to better fuse contextual information with visual features, ultimately improving its performance on downstream benchmarks.

4   Experimental Setup

In this section, we detail the model configurations and experimental settings used in our data augmentation, training, and experiments.

4.1   VLM architecture

We plug our C-ViT into two types of VLM architectures: encoder-decoder and decoder-only. For the encoder-decoder setup, we use Flan-T5 (Chung et al., 2024) as the backbone language model and evaluate performance on four VQA benchmarks: TextVQA, VQA-v2 (Antol et al., 2015), ST-VQA (Biten et al., 2019), and VizWiz (Gurari et al.,

2018). For the decoder-only architecture, we follow the LLaVA-1.5 configuration and use Vicuna-1.5 (Zheng et al., 2023) as the backbone. We evaluate our method on eight VQA benchmarks: TextVQA, VQA_{v2}, VizWiz, MME (Fu et al., 2023), POPE (Li et al., 2023b), MMBench (Liu et al., 2024c), ScienceQA (Lu et al., 2022) and SEED (Li et al., 2023a). For each architecture, we employ C-ViT as the vision encoder and compare its performance against two baselines: the original CLIP model (Radford et al., 2021) and QA-ViT (see Section 2).

4.2   Data Construction

We used some specific models in the process of building our CVIA dataset. We use MetaCLIP-h14 (Xu et al., 2024) as vision encoder to generate vision embeddings for similarity calculation. MiniCPM-V-2.6 (Yao et al., 2024) is used in both data augmentation pipeline for image filtering and new context (query) generation. For rewriting the visual inference, we adopt Llama-3.1-70B (Grattafiori et al., 2024).

4.3   Training Settings

For the contrastive learning stages (Stages 1 and 2), we use 1 million samples from our CVIA dataset for training, running for 10 epochs at each stage. The data distribution in $CVIA_{easy}$ consists of 40% basic, 40% easy, and 10% hard samples. In contrast, $CVIA_{hard}$ is composed of 10% basic, 10% easy, and 80% hard samples, emphasizing more challenging instances. For fine-tuning the encoder-decoder architecture, we follow the training strategy of QA-ViT and apply LoRA (Hu et al., 2022). We use the QA-ViT training data, excluding the OCR-VQA dataset (Mishra et al., 2019), due to partial unavailability. For the decoder-only architecture, we follow LLaVA-1.5, using the same data as in LLaVA's training process.

5   Results

Vision Encoder	TextVQA	VQA$_{v2}$	ST-VQA	VizWiz
CLIP-ViT	48.0	72.7	52.7	27.0
QA-ViT	49.7	73.2	54.5	27.6
C-ViT (ours)	**51.2**	**74.8**	**55.3**	**29.1**

Table 1: Evaluation results on Flan-T5-XL encoder-decoder architecture. The vision encoders in both QA-ViT and C-ViT are initialized with CLIP-ViT encoder weights, while their text encoders are initialized with Flan-T5-XL encoder weights.

Vision Encoder	TextVQA	VQA$_{v2}$	VizWiz	MME	POPE(rand/pop/adv)	MMBench	ScienceQA	SEED$_{img}$
CLIP-ViT	58.2	78.5	50.0	1510	87.3/86.1/84.2	64.3	66.8	66.1
QA-ViT	59.0	79.1	51.2	1518	88.0/87.2/85.3	65.1	67.0	66.2
C-ViT (ours)	**60.8**	**80.0**	**52.1**	**1533**	**89.1/87.6/86.1**	**66.2**	**68.9**	**67.2**

Table 2: Evaluation results on LLaVA-1.5 decoder-only architecture. All methods employ Vicuna-1.5-7B as the initial LLM to be trained with the vision encoder. The vision encoders in both QA-ViT and C-ViT are initialized with CLIP-ViT encoder weights, while their text encoders are initialized with the weights of CLIP's text encoder.

5.1 Evaluation on Downstream Tasks

We evaluate the effectiveness of our proposed C-ViT via plugging it into both encoder-decoder and decoder-only VLM architectures. We compare its performance against the late-fusion baseline and the QA-ViT method across a range of benchmarks.

For the encoder-decoder architecture, we adopt Flan-T5 (Chung et al., 2024) as the base model and evaluate on four datasets, as shown in Table 1. For the decoder-only architecture, we use Vicuna-1.5 as the base model and compare our approach to LLaVA-1.5 following LLaVA evaluation benchmarks, including eight benchmarks (see Table 2).

In both settings, our method consistently outperforms the baselines, demonstrating that our approach, which provides contextualized vision representations, enhances model performance across various VQA tasks. Furthermore, when compared to other VLMs of similar parameter sizes but with different base models, C-ViT surpasses the baselines on nearly all datasets (see Appendix C). Moreover, the modular and decoupled design of C-ViT enables it to be a general-purpose strategy for improving vision-language tasks. This means that C-ViT is independent of the specific backbone architecture and can be used with any visual encoder.

5.2 Generalization Ability on Zero-shot Datasets

Vision Encoder	M^{3}CoT	IllusionVQA	
		Comp	S-L
CLIP-ViT	27.1	30.8	24.6
QA-ViT	35.5	31.6	23.3
C-ViT (ours)	**38.2**	**31.7**	**27.9**

Table 3: Evaluation results on M^{3}CoT (Chen et al., 2024a) and IllusionVQA (Shahgir et al., 2024) using the same setup as in Table 2. In IllusionVQA, "Comp" and "S-L" denote the Comprehension and Soft-Localization tasks respectively.

We hypothesize that fusing contextual information directly into the vision encoder and training C-ViT via curriculum learning would yield more robust visual representations and, in turn, stronger overall VLM performance. To test this, we evaluate our approach in a zero-shot setting on two recent benchmarks: M^{3}CoT, which emphasizes the reasoning ability within the visual modality, and IllusionVQA, which assesses a model's behavior on absurd or contradictory images. Results (see Table 3) show that our method substantially outperforms the baselines, especially on M^{3}CoT dataset and the Soft-Localization (S-L) task in Illusion-VQA. It demonstrates that context-aware vision encoding confers strong generalization ability, responding well to OOD visual inputs.

6 Analysis

6.1 How curriculum learning effects the final performance?

Vision Encoder	TextVQA	VizWiz
CLIP-ViT	40.2	23.7
C-ViT	**45.1**	**26.2**
w/o curriculum (data)	44.4	24.4
w/o curriculum (freeze)	44.5	24.9
w/o curriculum (unfreeze)	44.7	25.5

Table 4: Accuracy on the TextVQA and VizWiz dataset with different contrastive learning strategy. The methods uses Flan-T5-base as the backbone language model.

Given a model and a dataset, there are often multiple strategies to adapt the model to a target task. In our experiments, we investigate the effect of curriculum learning and compare it against standard contrastive learning approaches (see Table 4).

We evaluated three different training configurations without curriculum strategy. In the w/o curriculum (**data**) setting, only samples labeled as hard in CVIA were used during training. In the w/o curriculum (**freeze**) setting, the text encoder parameters remained frozen throughout training. In contrast, for w/o curriculum (**unfreeze**), the text encoder was set to be fully trainable.

All three configurations lead to performance drops. The **data** setting causes the largest decline, suggesting that the absence of easier examples hin-

ders the model's ability to learn robust alignments. Both the **freeze** and **unfreeze** settings also negatively impact performance. We attribute this to different causes: freezing limits adaptation due to insufficient trainable parameters, while unfreezing risks diminishing pretrained knowledge through excessive parameter updates.

Overall, our experiments demonstrate that curriculum learning enables C-ViT to better balance the retention of pretrained knowledge with the adaptation to new tasks, resulting in improved performance on downstream evaluations.

6.2 Stratified Performance Analysis
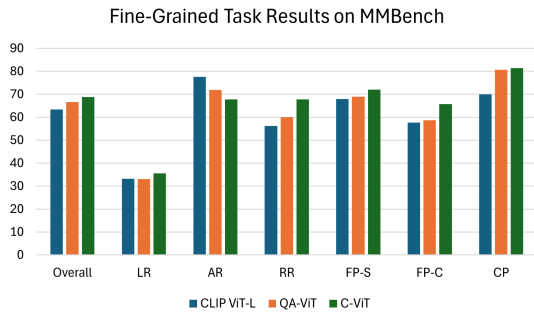
Figure 5: Comparison of accuracy on fine-grained tasks in MMBench. LR, AR, RR, FP-S, FP-C, and CP represent Logical Reasoning, Attribute Reasoning, Relation Reasoning, Fine-grained Perception (Single-instance), Fine-grained Perception (Cross-instance), and Coarse Perception, respectively.

In this section, we investigate deeper into the specific abilities that our C-ViT brings to a VLM. To this end, we evaluate C-ViT on MMBench, which is a benchmark that contains six subsets that evaluate on different aspects of the multimodal understanding. Figure 5 shows the detailed performance histogram across these subsets.

As illustrated in the figure, C-ViT achieves substantial improvements across all perception-related tasks, as well as in logical and relational reasoning tasks. These gains align with the core design of our context-aware approach. By introducing contextual information into the vision encoder at an early stage, the encoder is better equipped to produce semantically rich and context-relevant features, ultimately leading to improved performance.

However, an exception is observed in the attribute reasoning subset, where C-ViT demonstrates a drop in accuracy. Upon examining samples from this subset, we find that many of the prompts contain abstract or vague phrasing, such

as "The object shown in this figure:", which lacks meaningful semantic content. As a result, C-ViT struggles to effectively align the contextual information with the visual input. Therefore, this leads to the outcome that a well-trained C-ViT underperforms compared to the original CLIP.

6.3 Why do contextualized vision representations lead to accurate answers?

In this section, we investigate why VLMs tend to produce more accurate answers when using our C-ViT to generate contextualized vision representations. Our hypothesis is that C-ViT effectively emphasizes the relevant regions of an image that are critical for answering a given question or for supporting reasoning processes.

To investigate this hypothesis, we utilize Attention Rollout (Chefer et al., 2021) to visualize the vision encoder's attention through heatmaps. Attention Rollout computes a unified attention map by recursively aggregating attention matrices across all transformer layers. This process highlights the contribution of each input (image) token to the final output, offering insight into how contextual information guides visual reasoning and why our C-ViT performs better than all baselines.
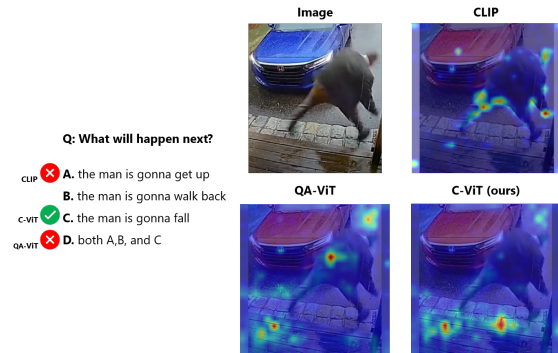

Figure 6: An example from MMBench. We compare attention heatmaps generated under the LLaVA-1.5 architecture using CLIP, QA-ViT, and C-ViT. In these heatmaps, the highlighted (red) regions indicate areas with greater influence on the model's output. Only our C-ViT setting selects the correct answer.

Figure 6 well explains that after fusing contextual information, our C-ViT significantly adjusts its self-attention patterns compared to context-agnostic image representations, even demonstrating a degree of visual reasoning ability. In this example, the main objects in the image are a man and a Honda car. Without specific contextual guidance, it is natural for CLIP to focus on the man's limbs,

7

and even the car logo. However, these regions are not relevant to the correct answer. QA-ViT, without being explicitly trained to fuse contextual information, focuses on the man's center of gravity and the wooden floor behind him. Although these features are somewhat related to the question, it fails to reach the correct answer. As a result, it concludes that all situations are equally plausible.

In contrast, our C-ViT demonstrates a markedly different attention pattern. Compared to the two baselines, the attention shifts significantly toward the wooden floor and the junction where it meets the stone road. This shift suggests that our method is aware of the height difference between the two surfaces. Remarkably, we can infer that the highlighted junction is the exact spot where the man's right leg previously missed the step. With C-ViT, the VLM focuses on the correct visual cues, enabling it to infer the reason for the man's posture and movement, and to make the correct deduction that he is about to fall.

This example provides strong evidence that contextualized vision representations offer more effective visual information to VLMs. By aligning attention with semantically relevant regions, they enhance the model's reasoning process and lead to more accurate answers in vision-language tasks.

7 Related Work

Contrastive Learning for Vision Models with Language. Contrastive learning (Hadsell et al., 2006) is a form of an unsupervised learning method that aims to learn data representation by maximizing the similarity between related samples and minimizing the similarity between unrelated samples.

Traditional vision models, such as convolutional neural networks (CNNs) (Szegedy et al., 2015; He et al., 2016) and Vision Transformers (ViTs) (Dosovitskiy et al., 2021), are typically limited to predicting predefined categories in classification tasks. Although many works (Chen et al., 2020; He et al., 2020) have attempted to enhance the capabilities of visual models through contrastive learning, CLIP (Radford et al., 2021) was the first to utilize it to align image and text modalities. CLIP consists of two parallel encoders for images and text. It learns a shared embedding space in which semantically aligned image-text pairs have high similarity scores. This multimodal alignment enables CLIP to generalize to novel visual concepts beyond the training categories (Mayilvahanan et al., 2024a).

In this work, we modified the CLIP framework so that the vision encoder can accept both vision and associated text inputs. Therefore, our contrastive learning objective is adapted to align the context-aware visual representation with its corresponding visual inference.

Enhanced Vision Representation. Many studies have explored strategies for improving visual representations (Li et al., 2021; Liu et al., 2023b; Vyskočil and Picek, 2023; Shi et al., 2024) to support more efficient image encoding. For example, BRAVE (Kar et al., 2024) generates its visual representation by employing k independent vision encoders and concatenating their outputs. This design enables the extraction of visual features from multiple perspectives, thereby enriching the visual representation space. However, the model may still struggle (Geigle et al., 2023) when image-text inputs require features beyond the collective scope of these encoders.

Chameleon (Team, 2024) adopts a different structure by treating images as discrete tokens like the text. It jointly embeds image and text tokens into an auto-regressive language model, allowing direct integration of visual and linguistic features. Despite its elegance, this method suffers from information loss during tokenization (Fan et al., 2024), as discrete tokens can only capture a limited subspace of the continuous visual feature space.

QA-ViT, as introduced in Section 2, enhances image representations by utilizing self-attention layers to integrate context information. However, it is only trained within the VLM architecture without individual training of vision encoding, limiting the model's performance, as discussed in Section 3.2.

8 Conclusion

We propose C-ViT, a contextualized vision transformer that fuse context information into the early-stage of visual encoding, enabling the model to construct dynamic context-aware visual representations. To support effective learning of such contextual representations, we introduce a fine-grained curriculum learning pipeline built upon our CVIA dataset. Extensive experiments across two VLM architectures demonstrate that C-ViT consistently improves reasoning robustness and generalization, particularly under OOD settings. Owing to its modular design, C-ViT can be integrated into a broad range of VLMs, offering a principled approach to enhancing VQA robustness in various scenarios.

Limitations

We acknowledge several limitations in our work. First, the dataset used for evaluation does not cover all domains, which may limit the ability of our method to specific or underrepresented fields. Second, both dataset construction and model evaluation use English as the main language, without consideration of multilingual scenarios. In addition, due to constraints in data and computational resources, we finetuned C-ViT using an existing vision encoder rather than training it from scratch, which may affect the optimality of the final model.

References

Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. 2024. Deciphering the role of representation disentanglement: Investigating compositional generalization in clip models. In *European Conference on Computer Vision*, pages 35–50. Springer.

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 IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluis Gomez, Marçal Rusiñol, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4290–4300. IEEE.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791. IEEE.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M³CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. 2024. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multimodal reasoning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13861–13871. IEEE Computer Society.

Gregor Geigle, Chen Liu, Jonas Pfeiffer, and Iryna Gurevych. 2023. One does not fit all! on the complementarity of vision encoders for vision and language tasks. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 97–117, Toronto, Canada. Association for Computational Linguistics.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *CoRR*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

9

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617. IEEE.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE Computer Society.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. 2024. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pages 113–132. Springer.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024a. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. 2024b. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Kai Liu, Zhihang Fu, Chao Chen, Sheng Jin, Ze Chen, Mingyuan Tao, Rongxin Jiang, and Jieping Ye. 2024b. Category-extensible out-of-distribution detection via hierarchical context descriptions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. 2023b. Mllms-augmented visual-language representation learning. *CoRR*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. 2024a. Does clip's generalization performance mainly stem from high train-test similarity? In *The Twelfth International Conference on Learning Representations*.

Prasanna Mayilvahanan, Roland S Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila Juhos, Matthias Bethge, and Wieland Brendel. 2024b. In search of forgotten domain generalization. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

10

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. 2024. Illusionvqa: A challenging optical illusion dataset for vision language models. In *First Conference on Language Modeling*.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, and 1 others. 2024. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *CoRR*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318. IEEE.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Jiří Vyskočil and Lukáš Picek. 2023. Vinvl+l: Enriching visual representation with location context in vqa. In *Computer Vision Winter Workshop*, CEUR Workshop Proceedings, Krems an der Donau, Austria.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. Demystifying clip data. In *The Twelfth International Conference on Learning Representations*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *CoRR*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2024. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *Transactions on Machine Learning Research*.

A Implementation Details

A.1 Model in Dataset Construction

Negative Image augmentation We utilize the full Cambrian dataset (10 million samples) to generate MetaCLIP embeddings and compute cosine similarity scores, which takes approximately 4 hours. For visual inference and filtering, we employ MiniCPM 2.6, requiring about 5 seconds to evaluate each negative image. Rewriting is performed using LLaMA-3.1 70B version, loaded in 4-bit precision, taking roughly 10 seconds per instance. All models are run on 1 NVIDIA A100-SXM4 (80GB) GPU.

Negative Context Augmentation We simultaneously load two independent MiniCPM 2.6 for ask and answer on 1 NVIDIA A100-SXM4 (80GB) GPU. On average, processing a single image takes about 20 seconds. The rewriting configuration remains the same as described above.

A.2 Model in Training

Contrastive Learning Stage In the contrastive learning stage, we fine-tune C-ViT using a batch size of 20 with the CLIP text encoder, and 80 with the encoder module of Flan-T5 XL. We use a learning rate of $1e^{-4}$, a cosine learning rate scheduler with warm-up step 300 and no weight decay. Training is based on 8 NVIDIA A100-SXM4 (80GB) GPUs, with a total training time of approximately 5 hours.

Supervised Finetuning Stage For encoder-decoder architecture, we finetune the VLM with a batch size 4 for 2 epochs, with a fixed random seed. We adopt the LoRA setting, with rank=128, α=256 and a dropout rate of 0.05, using a learning rate of $1e^{-4}$. The finetuning is also based on 8 NVIDIA A100-SXM4(80GB) GPUs, with training time about 20 hours.

For the decoder-only architecture, we largely follow the setup from LLaVA-1.5. The vision-language connector is pretrained with a total batch size of 256 and a learning rate of $1e^{-3}$, using a cosine learning rate scheduler with a warm-up ratio of 0.03 and with a fixed random seed. The full VLM is then fine-tuned with a total batch size of 128 and a learning rate of $2e^{-5}$, again using a cosine learning rate scheduler with a warm-up ratio of 0.03 and a weight decay of 0. The pretraining and finetuning conducted based on 8 NVIDIA A100-SXM4(80GB) GPUs, with training time about 6, 30 hours respectively.

B CVIA Construction Details

This section details the prompts used for data generation in our CVIA construction pipelines.

B.1 Visual Inference Generation and Filtering Prompts

Table 5, 6 provides examples of the prompt used in negative image augmentation pipeline to guide VLM in generating visual inference and filtering out those invalid inferences.

B.2 Ask-Answer Prompts

Table 7, 8 illustrates the prompt framework used for generating context in different types (questioner) and its correlated visual inference (answerer).

B.3 Rewrite Prompts

The prompt used in rewriting the visual inferences are shown in Table 9.

C Full Evaluation Results

We report the performance of C-ViT in comparison with other VLM architectures of similar parameter sizes, as shown in 10. Our method achieves the best performance on most datasets.

D Case Study

Although contextualized vision representations often help the model capture more effective features, they can sometimes misinterpret the context and extract misleading information. Take Fig. 7 as an example. While both vision encoders answer the question correctly, the attention heatmap of C-ViT reveals that it mistakenly focuses most on the inconspicuous declaration text at the bottom of the image. This occurs because C-ViT falsely identifies the text in the question as the key concept

and then associates it with the text in the image, leading it to believe that this text is relevant for visual inference. Despite this misinterpretation, C-ViT still produces the correct answer due to the partial retention of the ability from the original CLIP, which ensures that the model continues to allocate some attention to the sushi. This example highlights that C-ViT still requires improvement in performing implicit reasoning based on contextual information. At the same time, it also demonstrates C-ViT's strong ability to extract fine-grained image details.
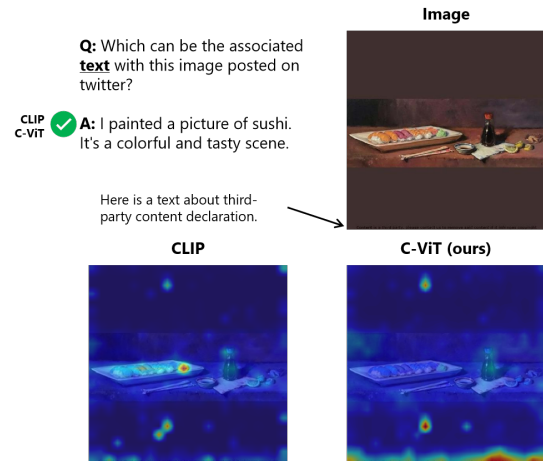

Figure 7: An example from MMBench. We compare attention heatmaps generated under the LLaVA-1.5 architecture using CLIP and C-ViT. In these heatmaps, highlighted (red) regions indicate areas with greater influence on the model's output. Although C-ViT answers correctly, it focuses on the wrong region

E Augmented Data Analysis

In this section, we present a analysis of the augmented data in CVIA.
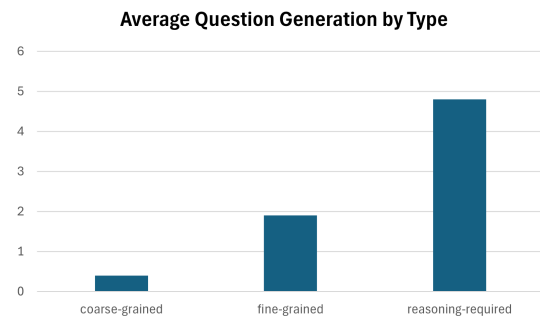
E.1 Negative Image Augmentation

Figure 8: The average number of questions by type generated by the context augmentation pipeline.

role	content
User	**Answer the question around** 5 **words. Do not hallucinate if not sure.** What kind of room am I looking at?
MiniCPM 2.6	Kitchen
User	**Is the answer correct given the question and image? Only answer yes or no, or no enough evidence. Question:** What kind of room am I looking at? **Answer:** You're looking at a kitchen.
MiniCPM 2.6	Yes.

Table 5: An example of our prompts. When the model's answer is "Yes," the image considered invalid.
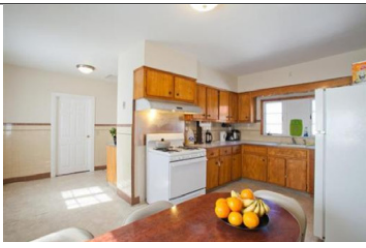
role	content
User	**Answer the question around** 5 **words. Do not hallucinate if not sure.** What kind of room am I looking at?
MiniCPM 2.6	Living room
User	**Is the answer correct given the question and image? Only answer yes or no, or no enough evidence. Question:** What kind of room am I looking at? **Answer:** You're looking at a kitchen.
MiniCPM 2.6	No.
User	**Is the answer correct given the question and image? Only answer yes or no, or no enough evidence. Question:** What kind of room am I looking at? **Answer:** You're looking at a living room.
MiniCPM 2.6	No.

Table 6: An example of our prompts. When the model answers two "No," the image is considered valid.

role	content
User	 **I have an image. Ask me questions about the content of this image. Each time ask one question only without giving an answer. Do not ask the same question as before. I'll put my answer beginning with "Answer:".**
Questioner (MiniCPM 2.6)	Describe this image in detail.
User	**Answer:** ... **Next Question. Avoid asking yes/no questions. Question should focus on border details.** **Question:**
Questioner	What is the color of the walls in this kitchen?
...	...
User	**Answer:** The kitchen's color scheme and furniture arrangement create a warm, inviting atmosphere. Light neutral walls complement wooden cabinetry, while beige counter tops add brightness without clashing with decorative elements. Strategic furniture placement enhances efficiency during meal prep, combining functionality with aesthetics to create a welcoming space. **Next Question. Avoid asking yes/no questions. Question should focus on reasoning.** **Question:**
Questioner	How does the presence of a fruit bowl in this kitchen setting contribute to its overall atmosphere and functionality?
...	...

Table 7: An example of our prompts used with the questioner. Notice that we force the questioner to use "Describe the image in detail" as the first question.

role	content
User	**Answer given questions. If you are not sure about the answer, say you doń know honestly. Don't imagine any contents that are not in the image.** **Question:** Describe this image in detail. **Answer:**
Answerer (MiniCPM 2.6)	The image shows a bright, functional kitchen with wooden cabinets, light-colored countertops, and neutral off-white walls. A white stove with a range hood and a refrigerator are neatly integrated, while small appliances like a coffee maker sit on the counters. In the foreground, a wooden dining table is adorned with a colorful fruit bowl. Natural light streams through a window, enhancing the clean and inviting atmosphere, complemented by simple ceiling lighting.
User	**Question:** What is the color of the walls in this kitchen? **Answer should be around** 48 **words. Answer:**
Answerer	The kitchen walls are painted a light, neutral color that complements the wooden cabinetry and beige countertops. This choice creates an inviting atmosphere, allowing furniture and decor to stand out without clashing, while also reflecting natural lighting to enhance brightness and make the space feel larger.

Table 8: An example of our prompts used with the answerer.

role	content
User	**Given the question, rewrite the answer to 5 words. question:** What kind of room am I looking at? **answer:** Kitchen
Llama 3.1	It's a kitchen room.

role	content
User	**Given the question, rewrite the answer to** 27 **words. question:** Are there any distinctive features or components you can identify on the bus? **answer:** The bus in the image has several distinctive features. It is a modern coach-style vehicle, characterized by its streamlined design and large windows that provide ample visibility for passengers inside. The number "3804" prominently displayed on it suggests this may be part of an organized fleet or public transportation system. Additionally, there are multiple doors visible along one side which likely serve as entry points to accommodate passenger boarding efficiently at stops with limited space like urban areas where parking might not allow buses full access from all sides.
Llama 3.1	The bus has a modern coach-style design, large windows, and the number "3804" displayed prominently, suggesting it's part of an organized fleet or public transportation system.

Table 9: Two examples of our prompts used to rewrite the visual inference.

Method	LLM	TextVQA	VQA$_{v2}$	POPE(rand/pop/adv)	MME
InstructBLIP	Vicuna-7B	50.1	-	-/-/-	1294
IDEFICS	LLaMA-7B	25.9	50.9	-/-/-	-
Qwen-VL	Qwen-7B	**63.8**	78.8	-/-/-	-
Qwen-VL-Chat	Qwen-7B	61.5	78.2	-/-/-	1488
LLaVA-1.5	Vicuna-7B	58.2	78.5	87.3/86.1/84/2	1511
LLaVA-1.5+C-ViT (ours)	Vicuna-7B	60.8	**80.0**	**89.1/87.6/86.1**	**1533**

Method	LLM	MMBench	VizWiz	ScienceQA	SEED$_{img}$
InstructBLIP	Vicuna-7B	36	34.5	60.5	58.8
IDEFICS	LLaMA-7B	48.2	35.5	-	44.5
Qwen-VL	Qwen-7B	38.2	35.2	67.1	62.3
Qwen-VL-Chat	Qwen-7B	60.6	38.9	68.2	65.4
LLaVA-1.5	Vicuna-7B	64.3	50.0	66.8	66.1
LLaVA-1.5+C-ViT (ours)	Vicuna-7B	**66.2**	**52.1**	**68.9**	**67.2**

Table 10: Evaluation results on eight VQA benchmarks. All baselines use CLIP-ViT as the vision encoder.

For the negative image component, each image in the Cambrian dataset finds 30 similar images on average above the predefined similarity threshold. However, only 15% of the positive images yield over five negative images above this threshold, which we used for data augmentation. After applying our filtering pipeline, approximately seven negative images per positive image are remained on average.

E.2 Negative Context Augmentation

The Cambrian dataset contains 3.9 million unique images. Since we designed a basic batch size of 10, each image must be associated with at least 10 distinct contexts. Our analysis shows that 12% of the images (approximately 470,000) already satisfy this requirement using existing data. For the remaining images, our pipeline generates on average 7.1 new contexts per image. The distribution of these generated contexts is depicted in Figure 8. Notably, the pipeline rarely generates contexts related to the image's coarse-grained features. Instead, it mainly produces reasoning-based contexts. It's because the majority of existing context belongs to the factual questions, both coarse-grained and fine-grained, making the model no longer capable of generating new factual questions. Overall, 70% of the images can obtain diverse contexts that meet our requirements.