

C³L: Content Correlated Vision-Language Instruction Tuning Data Generation via Contrastive Learning

Ji Ma^{1,2,3}, Wei Suo^{1,2,3}, Peng Wang^{1,2,3} and Yanning Zhang^{1,2,3}

¹School of Computer Science, Northwestern Polytechnical University, China

²Ningbo Institute, Northwestern Polytechnical University, China

³National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean, China
 {maji, suowei1994}@mail.nwpu.edu.cn, {peng.wang, ynzhang}@nwpu.edu.cn

Abstract

Vision-Language Instruction Tuning (VLIT) is a critical training phase for Large Vision-Language Models (LVLMs). With the improving capabilities of open-source LVLMs, researchers have increasingly turned to generate VLIT data by using open-source LVLMs and achieved significant progress. However, such data generation approaches are bottlenecked by the following challenges: 1) Since multi-modal models tend to be influenced by prior language knowledge, directly using LVLMs to generate VLIT data would inevitably lead to low content relevance between generated data and images. 2) To improve the ability of the models to generate VLIT data, previous methods have incorporated an additional training phase to boost the generative capacity. This process hurts the generalization of the models to unseen inputs (*i.e.*, “exposure bias” problem). In this paper, we propose a new Content Correlated VLIT data generation via Contrastive Learning (C³L). Specifically, we design a new content relevance module which enhances the content relevance between VLIT data and images by computing Image Instruction Correspondence Scores $S(I^2C)$. Moreover, a contrastive learning module is introduced to further boost the VLIT data generation capability of the LVLMs. A large number of automatic measures on four benchmarks show the effectiveness of our method.

1 Introduction

In recent years, significant advancements have been made in the field of natural language processing, with the emergence of Large Language Models (LLMs) revolutionizing the landscape of this field [Chung *et al.*, 2022; Chowdhery *et al.*, 2023; Touvron *et al.*, 2023]. Leveraging the powerful reasoning capabilities of LLMs, researchers [Li *et al.*, 2023c; Zhu *et al.*, 2023; Liu *et al.*, 2023c] have proposed integrating visual encoders with LLMs to construct multi-modal perception and reasoning systems. This integration empowers LLMs with the ability to perceive and process visual information, leading to the substantial strides of Large Vision-Language Models (LVLMs).

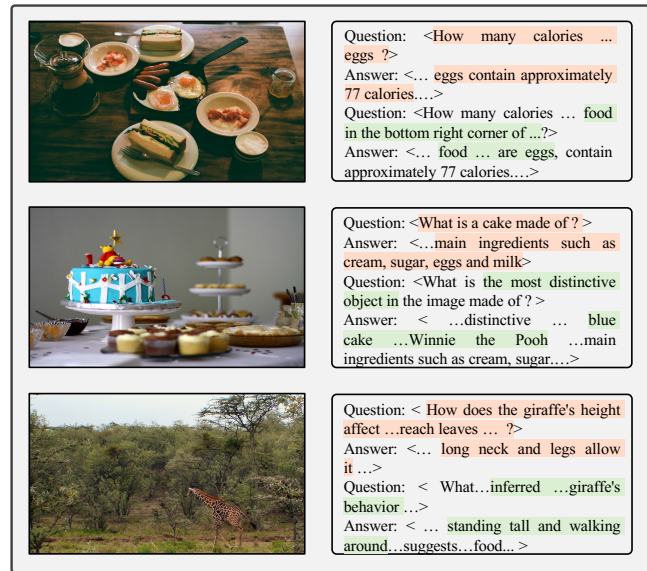


Figure 1: The illustration of the prior language knowledge problem when directly using current LVLMs to generate VLIT data. Existing models tend to generate data that exhibits low content relevance with the corresponding images (denoted in red). Our method effectively enhances the content relevance between VLIT data and images (denoted in green).

In practice, most of the existing LVLMs adopt a two-stage training paradigm [Liu *et al.*, 2023c; Zhu *et al.*, 2023; Liu *et al.*, 2023b; Li *et al.*, 2023c; Dai *et al.*, 2023]. In the first stage, a substantial amount of image-text pairs are used to pre-train LVLMs, typically employing the image-text contrastive approach [Liu *et al.*, 2023c; Liu *et al.*, 2023b; Zhu *et al.*, 2023]. The objective of this pre-training phase is to develop the fundamental cross-modal alignment capabilities of LVLMs.

In the second stage, a shift is made from traditional task-specific fine-tuning methods to a more general approach [Li *et al.*, 2023b]. Instead of relying on task-specific data for individual downstream tasks [Devlin *et al.*, 2018; Raffel *et al.*, 2020], models are now fine-tuned using high-quality Vision-Language Instruction Tuning (VLIT) data. This approach aims to develop general abilities of the models to under-

stand and follow various types of instructions while generating helpful, factual, and harmless responses. The primary approach [Liu *et al.*, 2023c; Wang *et al.*, 2023b] to generating VLIT data is through using GPT-4(V) [OpenAI, 2023] to reduce manual annotations. With the improving capabilities of open-source LVLMs, researchers are increasingly applying these models to generate VLIT data, enabling them to overcome the limited accessibility of GPT-4(V) [Wang *et al.*, 2023a; Bai *et al.*, 2023; Li *et al.*, 2023b].

Although significant progress has been made, this paradigm still faces the following challenges: 1) A longstanding issue with multi-modal models is to overly depend on prior language knowledge [Goyal *et al.*, 2017], leading them to ignore visual content. Therefore, when generating VLIT data using LVLMs, the models often fail to effectively focus on the visual information and instead heavily rely on prior language knowledge. As shown in Fig. 1, the resulting VLIT data (denoted in red) demonstrates limited correlation with the corresponding images. 2) Current LVLMs are primarily trained with an emphasis on developing strong reasoning capabilities rather than their abilities as data generators. In fact, a straightforward approach [Wang *et al.*, 2023a] to address this issue is by introducing an additional training phase where the models are trained with high-quality VLIT data as ground truth labels. As evidenced in [Ranzato *et al.*, 2015; Lee *et al.*, 2020], models that are only exposed to correct samples suffer from the “exposure bias” problem, which will restrict generalization abilities of the models when encountering unseen samples. Therefore, this training paradigm hinders the effectiveness of turning the models into data generators.

To address the aforementioned challenges, we propose a new Content Correlated VLIT data generation via Contrastive Learning, which is called C^3L for short. In C^3L , we first apply the LVLM to generate a set of initial VLIT data. Then, to enhance the content relevance between VLIT data and images, we propose a novel content relevance module where Image Instruction Correspondence (I^2C) scores are computed based on whether or not the images are provided. On the other hand, due to the training paradigm that only utilizes high-quality VLIT data as ground truth labels lacks exposure to low-quality samples, we divide the initial VLIT data into two categories based on the I^2C scores. The data samples with high I^2C scores are considered as positive pseudo-labels, and vice versa. By employing contrastive learning framework [Lee *et al.*, 2020] with positive-negative pseudo-labels, we can effectively alleviate the “exposure bias” problem and enhance the capability of the model as a data generator. Benefiting from the above methods, C^3L effectively improves the content relevance of the model-generated VLIT data with images and enhances the capability of the model as a data generator.

According to automatic evaluations, fine-tuned LVLMs using data generated by C^3L achieve comparable performance to the state-of-the-art models on four recent multi-modal benchmarks. More importantly, only 5k VLIT data samples generated by our C^3L are used to fine-tune these models, significantly reduces the computational cost. In summary, we make the following contributions:

1) We propose a new content relevance module that can model the content correlation between VLIT data and images. This module effectively improves the content relevance and the utilization of visual information.

2) We develop an advanced contrastive learning module for VLIT data generation, which applies generated samples as pseudo-labels to boost the data generation capacity of the LVLMs further. To the best of our knowledge, we are the first to explore contrastive learning on the VLIT data generation.

3) The LVLMs fine-tuned using data generated by our C^3L achieves comparable or even better performance on four multi-modal benchmarks (*i.e.*, SEED [Li *et al.*, 2023a], LLaVA^W [Liu *et al.*, 2023b], MMB [Liu *et al.*, 2023d] and POPE [Li *et al.*, 2023d]). Meanwhile, automatic measures and ablation studies all show the effectiveness of our method.

2 Related Works

2.1 Large Vision-Language Models

Most current LVLMs typically comprise three main components [Hamadi, 2023; Gu *et al.*, 2023]: visual encoder, LLM and projection layer to connect visual encoder and LLM. In order to connect the visual encoder and the LLM, a connectivity structure is employed [Dai *et al.*, 2023; Liu *et al.*, 2023c; Zhu *et al.*, 2023]. Typically, the majority of LVLMs adopt a two-stage training paradigm, encompassing Vision-Language Pre-training (VLP) and VLIT. In the first stage, large scale image-text pairs are employed [Li *et al.*, 2023c; Gan *et al.*, 2022] to establish fundamental vision-language alignment of LVLMs. This process involves training models to comprehend the visual information in the images and generate captions that accurately depict the visual content. In the second stage, a significant corpus of instruction data generated by models [Bai *et al.*, 2023; Wang *et al.*, 2023b] is utilized to enhance the capacity of LVLMs in comprehensively understanding vision-language instructions and generating appropriate responses. Existing LVLMs tend to excessively rely on prior language knowledge [Guan *et al.*, 2023]. Therefore, when employing them to generate VLIT data, the generated data exhibits limited content relevance with images. In order to alleviate this issue, we design a novel content relevance module to improve the content correlation between model-generated VLIT data and corresponding images.

2.2 VLIT Data Generation

In order to enhance the capabilities of LVLMs to understand and follow instructions, the VLIT phase based on VLIT data has been introduced [Li *et al.*, 2023b; Hamadi, 2023]. To reduce the cost of human annotation, the generation of VLIT data primarily relies on automatic model generation [Li *et al.*, 2023b]. This is achieved by utilizing closed-source, powerful large models such as GPT-4(V) [Liu *et al.*, 2023c; Wang *et al.*, 2023b]. Specifically, [Liu *et al.*, 2023c] collects VLIT data based on the existing image-pair datasets. By prompting text-only GPT-4 [OpenAI, 2023], high-quality question-answer pairs are obtained and used as VLIT data (*i.e.*, LLaVA-158k). Moreover, as generating VLIT data solely by prompting text-only models like GPT-4 would

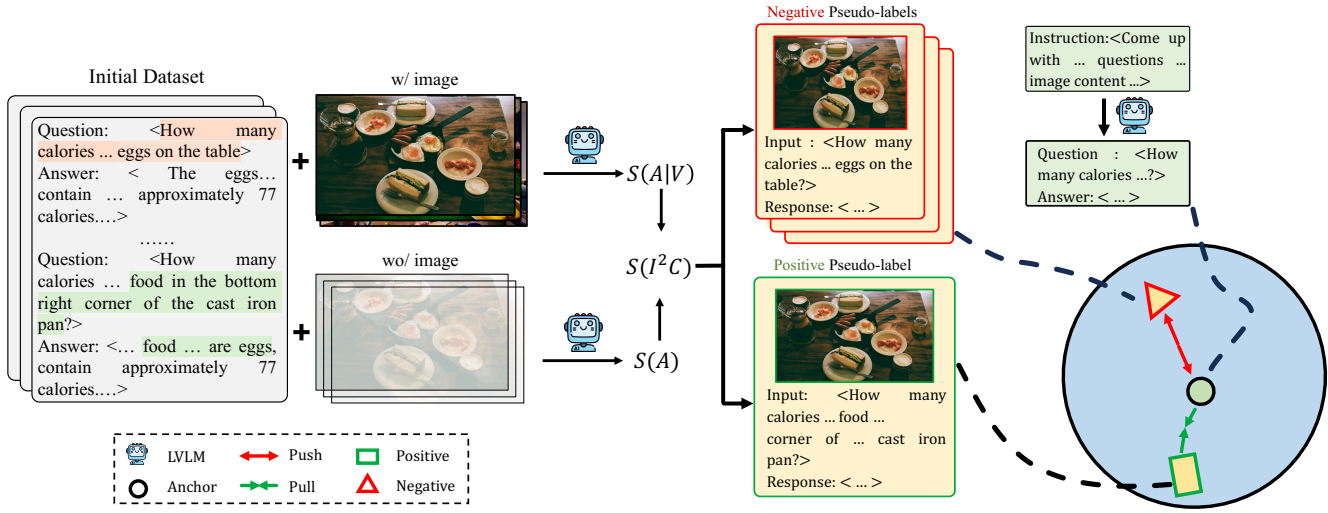


Figure 2: Overview of our Content Correlated VLIT data generation via Contrastive Learning (C^3L). Given the initial dataset and corresponding images, we first use Content Relevance module to obtain the I^2C scores based on whether or not the image is provided. Then, positive-negative pseudo-labels are selected based on I^2C scores. Further, our Contrastive Learning module maximize the similarity between the anchor and positive pseudo-label while minimizing the similarity between the anchor and negative pseudo-labels.

inevitably result in the loss of visual information, previous work [Wang *et al.*, 2023b] proposes to leverage GPT-4V [OpenAI, 2023] to generate VLIT data with the entire visual context. Recently, as open-source LVLms continue to evolve, researchers [Bai *et al.*, 2023; Wang *et al.*, 2023a] have turned to generate VLIT data by LVLms. [Bai *et al.*, 2023] employs self-instruction to acquire VLIT data, which can be used to improve the image content comprehension ability of the LVLms. These approaches contribute to further reducing the cost associated with assessing GPT-4(V). However, as current open-source LVLms are primarily trained with an emphasis on developing strong reasoning abilities, directly applying the models to generate VLIT data may result in undesirable results. Different from the above methods, we introduce a new contrastive learning module to boost the data generation capacity of the LVLms.

2.3 Contrastive Learning

Contrastive learning proposed in [Hadsell *et al.*, 2006] has been demonstrated to be an effective method for visual feature extraction [He *et al.*, 2020]. [Chen *et al.*, 2020] shows that contrastive learning benefits from larger batch sizes and can boost the performance of self-supervised learning in computer vision tasks. Some studies have also applied contrastive learning to multi-modal sequence modeling tasks, such as image captioning [Dai and Lin, 2017] and visual question answering [Liang *et al.*, 2020; Lai *et al.*, 2023; Suo *et al.*, 2023]. However, these studies primarily focus on utilizing contrastive learning methods to improve the multi-modal reasoning abilities and robustness of models. In contrast, our contrastive learning module is applied to enhance the capability of the models in generating VLIT data that exhibits higher content relevance with images. By employing positive-negative pseudo-labels with the contrastive learning framework, our contrastive learning module can alleviate the

“exposure bias” problem [Ranzato *et al.*, 2015] which hurts the generalization of the models and thereby transform the model into a more capable data generator.

3 Method

The aim of this paper is to generate VLIT data that exhibits high content relevance with corresponding images and enhance the capability of the model as a data generator. We achieve this by addressing two issues that prior works have ignored: 1) Due to the tendency of multi-modal models to overly rely on language prior knowledge [Goyal *et al.*, 2017], the relevance between the model-generated data and images is limited. 2) Considering the low capacity of current LVLms in generating data, VIGC [Wang *et al.*, 2023a] introduces an additional training phase to transform the model into a data generator using high-quality VLIT data as ground truth labels. This training paradigm has been demonstrated to result in the “exposure bias” problem [Ranzato *et al.*, 2015; Lee *et al.*, 2020]. In this section, we introduce our Content Correlated VLIT data generation via Contrastive Learning (C^3L). As shown in Fig. 2, the C^3L comprises an augmented pipeline with content relevance module and contrastive learning module. Next, we would introduce our method in detail.

3.1 Conventional VLIT Data Generation Pipeline

VLIT data. Given the image I_i , the VLIT data is structured as question-answer pairs $[Q_j, A_j]$, where $j \in \{1, 2, \dots, N_j\}$ denotes the j -th question-answer pair associated with the image I_i .

Conventional pipeline. As in [Wang *et al.*, 2023a], the LVLm that used for generating VLIT data is first trained with high-quality VLIT data as ground truth labels. Next, given a set of instructions and images, the enhanced LVLm is utilized to generate VLIT data which is in the form of question-

answer pairs $[Q_j, A_j]$. The instructions here are used to guide the enhanced LVLM in generating VLIT data. Finally, the data can be used to fine-tune LVLMs in order to enhance the general capabilities of the models.

Limitations. 1) The process overlooks the impact of language priors [Goyal *et al.*, 2017] embedded in LLM counterparts of LVLMs, resulting in limited correlation between the generated VLIT data and the corresponding images. 2) The training phase only uses high-quality VLIT data as ground truth labels. As evidenced in [Lee *et al.*, 2020], this training paradigm suffers from the “exposure bias” problem [Ranzato *et al.*, 2015] and hinders the effectiveness of transforming the model into a data generator.

3.2 Content Relevance Module

To solve the first issue, we propose a novel content relevance module which computes the **Image Instruction Correspondence Scores** $S(I^2C)$ based on whether the image is present or not.

Given a set of instructions and images, we first apply a widely used LVLM such as LLaVA [Liu *et al.*, 2023c] to generate the initial VLIT data. Due to the influence of language prior knowledge, the initial dataset contains data samples that exhibit low relevance to the image content, as shown in Fig. 1. Next, for a given data sample $[Q_j, A_j]$ in our initial VLIT dataset, we set Q_j and the corresponding image as inputs to the model. After obtaining the outputs of the model, we compute the probability value p_t^v of the outputs for each answer token in A_j , where t denotes the t -th token. Then, these probability values are concatenated to obtain the Visual Answer Scores $S(A|V)$. The $S(A|V)$ scores measure the responses of the LVLM to a specific VLIT data given the image. We can also compute the probability values p_t^d and obtain the Direct Answer Scores $S(A)$ in the absence of the image. The $S(A)$ scores quantify the responses of the model to a given VLIT data in the absence of an image. The final scores $S(I^2C)$ is computed by using KL-divergence:

$$\begin{aligned} S(I^2C) &= D_{KL}(S(A|V)||S(A)) \\ &= \sum_{t=1}^n p_t^v \cdot \log \frac{p_t^v}{p_t^d}. \end{aligned} \quad (1)$$

The $S(I^2C)$ scores measure the differences in responses of the model when provided with and without an image. Based on the properties of KL-divergence [Bu *et al.*, 2018], a low $S(I^2C)$ means that the divergence between $S(A|V)$ and $S(A)$ is small, indicating that the content relevance between the data sample and the image is low.

After obtaining the $S(I^2C)$ scores for all initial VLIT data samples, we train the model by using 10 % data samples with higher $S(I^2C)$ as ground truth labels, the standard cross-entropy loss L_r is used during training.

3.3 Contrastive Learning Module

Due to the presence of the “exposure bias” problem [Ranzato *et al.*, 2015] when training models solely with high-quality VLIT data as ground truth labels, we introduce an additional contrastive learning module in which the model also learns

experience from low-quality data samples in the initial VLIT dataset. Following contrastive learning framework [Lee *et al.*, 2020], we maximize the similarity between the anchor and positive pseudo-labels while minimizing the similarity between the anchor and negative pseudo-labels as follows.

In particular, given an image I_i , the data sample $[Q_j, A_j]$ with the highest $S(I^2C)$ is selected and employed as the positive pseudo-label. We concatenate $[Q_j, A_j]$ and embed the sequence s using text encoder in the LVLM, denoted as $\mathbf{e}_s \in \mathbb{R}^{d \times l}$ where l represents the length of sequence. Affine transformation ξ with the ReLU and AvgPool is used to project \mathbf{e}_s onto the latent space $\mathbf{h}_s \in \mathbb{R}^d$. This process is denoted as:

$$\begin{aligned} \mathbf{e}_s &= \text{Embedding}(Q_j, A_j), \\ \mathbf{h}_s &= \xi(\mathbf{e}_s), \end{aligned} \quad (2)$$

where $\xi(\cdot) = \text{AvgPool}(\text{ReLU}(\cdot))$.

Similarly, the remaining data samples $\hat{s} \in \mathbf{S}$ corresponding to the image I_i are selected and employed as the negative pseudo-labels. The representations $\mathbf{h}_{\hat{s}}$ are computed using eq.2. Then, we input an instruction along with the image I_i into the model and obtain the anchor data y . Finally, we compute the contrastive loss L_c as follows:

$$L_c = -\log \frac{\exp(\text{sim}(\mathbf{h}_s, \mathbf{h}_y)/\tau)}{\sum_{\hat{s} \in \mathbf{S}} \exp(\text{sim}(\mathbf{h}_{\hat{s}}, \mathbf{h}_y)/\tau)}, \quad (3)$$

where $\text{sim}(\cdot, \cdot)$ is a cosine similarity function.

3.4 Augmented VLIT Data Generation Pipeline

By using our content relevance module and contrastive learning module, we can construct an augmented VLIT data generation pipeline.

Initial data generation. Given a set of instructions and images, we first employ a widely used LVLM such as LLaVA [Liu *et al.*, 2023c] to generate a set of initial VLIT data, wherein the data is in the form of question-answer pairs. Inspired by [Changpinyo *et al.*, 2022], we use a simple two-step generation (*i.e.*, first generate captions for the images and then use the LLM-counterparts of LVLMs to generate data based on captions) to further fill the gap between the modalities. Following [Wang *et al.*, 2023a], the instructions are primarily categorized into three main classes: 1) conversation 2) detailed description 3) complex reasoning. The instructions are used to guide the model in generating VLIT data (*e.g.*, *Generate five in-depth reasoning questions and then answer them based on the image.* or *Generate five questions in order to describe the image and then answer them.*).

Content relevance module. Due to the language priors embedded in LLM counterparts of LVLMs, the initial VLIT dataset contains data samples that exhibit low content relevance between the VLIT data and the corresponding images. By using our content relevance module, we can compute the $S(I^2C)$ for all VLIT data samples based on whether or not the image is provided and train the model by using 10 % data samples with higher $S(I^2C)$ as ground truth labels. The standard cross-entropy loss L_r is used during this training phase.

Method	Data	Base LLM	SEED	LLaVA ^W	MMB	POPE
MiniGPT-4 [Zhu <i>et al.</i> , 2023]	3.5k	Vicuna-7b	43.3	-	24.3	74.1
VisualGLM [Du <i>et al.</i> , 2022]	-	ChatGLM-6b	47.0	-	37.6	-
LLaVA [Liu <i>et al.</i> , 2023c]	158k	Vicuna-7b	49.4	-	49.9	67.8
BLIP2 [Li <i>et al.</i> , 2023c]	-	Vicuna-13b	46.4	38.1	-	85.3
InstructBlip [Dai <i>et al.</i> , 2023]	1.2M	Vicuna-7b	53.4	60.9	36.0	83.7
InstructBlip [Dai <i>et al.</i> , 2023]	1.2M	Vicuna-13b	-	58.2	-	78.9
IDEFICS-80B [Laurençon <i>et al.</i> , 2023]	1M	Llama-65b	52.0	-	54.6	-
MiniGPT-4 w/ C^3L^*	5k	Llama-7b	44.1	59.7	42.6	70.3
MiniGPT-4 w/ C^3L	5k	Llama-7b	40.9	63.6	37.0	70.4
LLaVA w/ C^3L^*	5k	Vicuna-7b	52.7	62.2	47.6	75.7
LLaVA w/ C^3L	5k	Vicuna-7b	56.1	61.7	51.9	76.0

Table 1: Comparison with the state-of-the-art methods on four benchmarks. C^3L^* denotes the LVLM is fine-tuned using VLIT data generated by the other model (e.g., MiniGPT-4 w/ C^3L^* denotes that data is generated by LLaVA).

Dataset	Instances	Avg. Q len	Avg. A len
ComVint [Du <i>et al.</i> , 2023a]	32k	23.3	18.9
LLaVA-158k [Liu <i>et al.</i> , 2023c]	158k	11.0	66.7
LRV [Liu <i>et al.</i> , 2023a]	400k	12.1	15.2
Ours	5k	23.4	12.8

Table 2: Statistic of VLIT Dataset. “Instances” represents the total number of data instances in the dataset. “Avg. Q len” denotes the average length of questions in the dataset, while “Avg. A len” denotes the average length of answers.

Contrastive learning module. Due to the presence of “exposure bias” problem, we introduce the contrastive learning module for models to learn experience from low-quality samples in the initial dataset. Specifically, given an image I_i , the data sample with the highest $S(I^2C)$ is employed as the positive pseudo-label and the remaining data samples as negative pseudo-labels. By incorporating the positive-negative pseudo-labels and the contrastive learning framework, our module effectively enhances the capability of the LVLM to generate VLIT data.

Final data generation. The final 5k VLIT data is generated using the enhanced LVLM along with the two-step generation and can be employed to fine-tune other LVLMs, thereby enhancing the overall capacity of the models.

4 Experiments

4.1 Experimental Setting

Implementation details. We select two representative LVLMs (i.e., LLaVA [Liu *et al.*, 2023b] and MiniGPT-4 [Zhu *et al.*, 2023]) as our backbones to demonstrate the generality of our method. During the initial data generation phase, we utilize images from the COCO 2014 [Lin *et al.*, 2014] training set and generate 20k initial VLIT data samples. When fine-tuning backbones using our final data, following [Liu *et al.*, 2023c], we fix the weights of the CLIP [Radford *et al.*, 2021] visual encoder during fine-tuning. The AdamW [Kingma and

Dataset	Model	SEED
ComVint [Du <i>et al.</i> , 2023a]	LLaVA-7B	54.21
LLaVA-158k [Liu <i>et al.</i> , 2023c]	LLaVA-7B	49.53
LRV [Liu <i>et al.</i> , 2023a]	LLaVA-7B	56.58
Ours	LLaVA-7B	56.09

Table 3: The evaluation results on SEED after fine-tuning LLaVA-7B using different VLIT datasets.

Ba, 2014] is used as our optimizer with the weight decay $1e-5$. We fine-tune all models with a learning rate of $2e-5$ and batch size is set to 2. For MiniGPT-4, two 3090Ti GPUs are used while for LLaVA we use eight 2080Ti GPUs.

Benchmarks. To comprehensively evaluate the performance of LVLMs after VLIT using data generated by our C^3L , we select four benchmarks (i.e., SEED, MMB, LLaVA^W and POPE) for evaluation.

1) SEED [Li *et al.*, 2023a] consists of 19k multiple-choice questions to evaluate the image and video understanding capabilities of LVLMs across 12 different tasks. Following [Du *et al.*, 2023b], we only evaluate the capacity of LVLMs on image-related tasks.

2) MMB [Liu *et al.*, 2023d] develops a comprehensive evaluation pipeline which incorporates a novel *CircularEval* strategy to effectively evaluate the capabilities of LVLMs systematically. We use *dev* split of MMB for evaluation.

3) LLaVA^W [Liu *et al.*, 2023b] encompasses indoor and outdoor scenes, memes, paintings, sketches, etc, to evaluate the capabilities of LVLMs in more challenging tasks.

4) POPE [Li *et al.*, 2023d] aims to effectively evaluate the common issue of hallucinations in almost all LVLMs. By constructing three types of sampling strategies along with a polling-based query method, they can accurately measure the severity of model hallucinations for different objects.

Generated dataset. As shown in Table 2, we conduct a statistical analysis of our final VLIT dataset. It is evident that

	instructions	filtering	CRM	CLM	SEED	POPE
1	✓	✓	-	-	53.2	73.2
2	✓	✓	-	✓	54.1	74.4
3	✓	✓	✓	-	55.8	75.1
4	✓	✓	✓	✓	56.1	76.0

Table 4: **Ablation study.** We ablate key components to demonstrate the effectiveness of our method. The “instructions” denotes the instructions used to guide LVLMs to generate data. The “filtering” represents the basic filtering process that removes duplicate or invalid data identified by heuristics (*e.g.*, data is too short or too long). “CRM” and “CLM” are Content Relevance Module and Contrastive Learning Module respectively.

Method	SEED
LLaVA-7B w/ Self-Instruct [Wang <i>et al.</i> , 2022]	53.9
LLaVA-7B w/ VIGC [Wang <i>et al.</i> , 2023a]	55.2
LLaVA-7B w/ I^2C (Ours)	56.1
LLaVA-13B w/ Self-Instruct [Wang <i>et al.</i> , 2022]	56.6
LLaVA-13B w/ VIGC [Wang <i>et al.</i> , 2023a]	57.4
LLaVA-13B w/ I^2C (Ours)	58.3

Table 5: **Alternative data selection methods testing.** We compare with two other data selection methods on SEED to test the effectiveness of our I^2C scores.

our dataset of 5k instances is smaller in number compared to other VLIT datasets, such as LRV-400k [Liu *et al.*, 2023a]. In Table 3, it can be observed that the performance of LLaVA-7B [Liu *et al.*, 2023c] when fine-tuned using our 5k data is competitive to the performance achieved when using LRV-400k, suggesting the effectiveness of our C^3L .

4.2 Quantitative Evaluation

To assess the improvements in the overall capabilities of LVLMs fine-tuned using data generated by our C^3L , we compare these fine-tuned LVLMs with the state-of-the-arts models on four benchmarks in Table 1. The C^3L^* denotes that VLIT data used to fine-tune the LVLm is generated by the other LVLm (*e.g.*, MiniGPT-4 w/ C^3L^* denotes that data is generated by LLaVA). We observe that LVLMs with C^3L achieve comparable or even better results on these four widely used benchmarks compared to other state-of-the-art models. More importantly, only 5k VLIT data generated by models is utilized to fine-tune them. On the SEED, LLaVA w/ C^3L outperforms the previous models with an improvement of 2.7%. This demonstrates the effectiveness of our C^3L in improving the content relevance between VLIT data and images, thereby further enhancing the image understanding capabilities of LVLMs. LLaVA^W is designed to evaluate the capacity of LVLMs in more challenging tasks, we observe that MiniGPT-4 w/ C^3L achieves better performance than previous models on this benchmark. On the POPE, models with C^3L achieve comparable results with previous models which suggests that data generated by our C^3L exhibits higher content relevance with images, thus can be applied to improve the factualness in LVLMs.

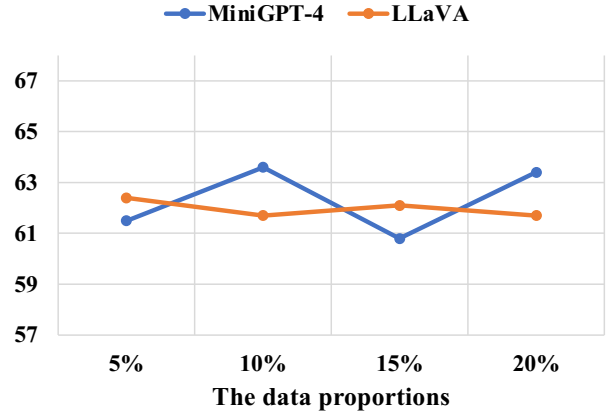


Figure 3: **Alternative data selection proportions testing.** We conduct experiments on LLaVA^W to test the effects of different data selection proportions.

4.3 Ablation Studies

In order to demonstrate the effectiveness of our method in improving the overall capacity of the LVLm including factualness, we conduct several ablation studies on the SEED and POPE. As shown in Table 4, in the first row, the baseline is established by fine-tuning LLaVA-7B [Liu *et al.*, 2023c] using VLIT data generated with instructions and basic filtering. In the second row, it can be found that the performance on SEED and POPE exhibits an improvement of 0.9% and 1.2% respectively. This could be attributed to the effectiveness of our contrastive learning module in solving the “exposure bias” problem and enhancing the capacity of the model as a capable data generator. In the third row, we can find that our content relevance module plays a pivotal role, bringing a 2.6% improvement on SEED and a 1.9% improvement on POPE in comparison to the baseline model. These results prove that our content relevance module can enhance the correspondence between VLIT data and images, thus improving the overall capabilities of models when fine-tuned using the data. In the last row, it can be observed that additional improvements are achieved when the two modules work collaboratively, which demonstrate that our C^3L is remarkably effective in improving the content relevance of the model-generated VLIT data with images and enhancing the capability of the model as a data generator.

4.4 Alternative Data Selection Setting

We propose two alternative data selection settings and conduct several experiments to measure the reliability of our method.

Alternative data selection methods. To further demonstrate the effectiveness of our $S(I^2C)$ in selecting positive/negative pseudo-labels, we compare our method with two other approaches (*i.e.*, Self-Instruct [Wang *et al.*, 2022] and VIGC [Wang *et al.*, 2023a]) on SEED. As shown in Table 5, we conduct experiments by replacing the I^2C scores with these two methods using two backbones (*i.e.*, LLaVA-7B and LLaVA-13B). For Self-Instruct, we apply the same

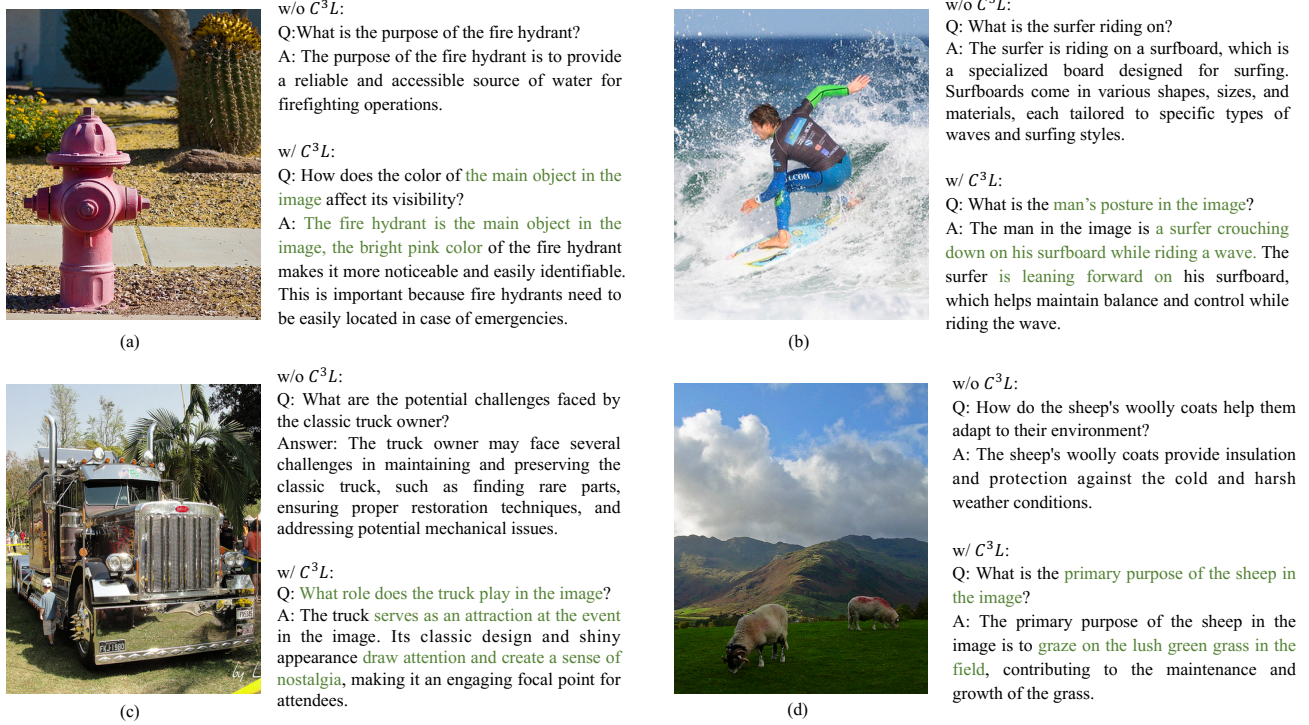


Figure 4: **Generation results.** We show the VLIT data generated w/o C^3L and w/ C^3L .

filtering and post-processing methods following [Wang *et al.*, 2022] to identify low-quality data, the remaining instruction samples are considered as positive pseudo-labels. For VIGC, we generate data using two approaches: direct generation and iterative generation [Wang *et al.*, 2023a]. The data generated by iterative generation is considered as positive pseudo-labels. The results show that selecting positive/negative pseudo-labels by $S(I^2C)$ is more effective compared to other alternative data selection methods.

Alternative data selection proportions. In order to investigate the impact of the data proportions selected by content relevance module on the performance of the LVLMS (*i.e.*, LLaVA and MiniGPT-4) fine-tuned with C^3L , we compare the performance of LVLMS on LLaVA^W after fine-tuning with different data proportions. As shown in Fig. 3, the blue line denotes the performance of MiniGPT-4 changes with different data selection proportions, while the orange line corresponds to LLaVA. The results demonstrate that there is a marginal difference in the performance of LVLMS when fine-tuned with different data proportions. Therefore, we choose 10% data in the content relevance module in order to balance performance and cost.

4.5 Qualitative Evaluation

As shown in Fig. 4, we show some qualitative results generated by C^3L . Through overall comparison, the data generated by our C^3L exhibits higher content relevance with images. For example, in Fig. 4 (a), answering the question generated without C^3L does not require leveraging information from

the image, as the fire hydrant is primarily used to supply water during fire incidents. On the contrary, answering the question generated by C^3L not only requires common knowledge about the fire hydrant but also necessitates a comprehensive understanding of the image. Additionally, the data generated by C^3L is also more detailed and factual. In Fig. 4 (b)-(d), the generated data containing the question about the man's posture and the answer about the role of the truck, which is "serves as an attraction at the event".

5 Conclusion

In this paper, we propose a new Content Correlated Vision-Language Instruction Tuning Data Generation via Contrastive Learning (C^3L). We first design a new content relevance module to improve the correlation between VLIT data and images. Meanwhile, we propose a contrastive learning module, which utilizes the generated samples as pseudo-labels to boost the data generation capacity of the LVLMS. Furthermore, an augmented VLIT data generation pipeline is proposed by combining these two modules, which can be applied to generate VLIT data that exhibits high content relevance with images. According to automatic evaluations, fine-tuned LVLMS using data generated by our C^3L achieve comparable performance to the state-of-the-art models on four benchmarks and provides a new paradigm for our community.

Contribution Statement

Ji Ma and Wei Suo contribute equally to this work. Peng Wang is the corresponding author.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.U23B2013), Shaanxi Provincial Key R&D Program (No.2021KWZ-03), and Natural Science Basic Research Program of Shaanxi (No.2021JCW-03).

References

- [Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [Bu *et al.*, 2018] Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of kl divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674, 2018.
- [Changpinyo *et al.*, 2022] Soravit Changpinyo, Doron Kukliansky, Idan Szepes, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions, 2022.
- [Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.
- [Chowdhery *et al.*, 2023] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- [Chung *et al.*, 2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [Dai and Lin, 2017] Bo Dai and Dahua Lin. Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*, 30, 2017.
- [Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- [Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [Du *et al.*, 2022] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.
- [Du *et al.*, 2023a] Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv preprint arXiv:2311.01487*, 2023.
- [Du *et al.*, 2023b] Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning, 2023.
- [Gan *et al.*, 2022] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.
- [Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- [Gu *et al.*, 2023] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.
- [Guan *et al.*, 2023] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and Dinesh Manocha Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pages arXiv–2310, 2023.
- [Hadsell *et al.*, 2006] Raia Hadsell, Sumit Chopra, and Yann LeCun. 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, 2006.
- [Hamadi, 2023] Raby Hamadi. Large language models meet computer vision: A brief survey. *arXiv preprint arXiv:2311.16673*, 2023.
- [He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [Lai *et al.*, 2023] Chengen Lai, Shengli Song, Shiqi Meng, Jingyang Li, Sitong Yan, and Guangneng Hu. Towards more faithful natural language explanation using multi-level contrastive learning in vqa. *arXiv preprint arXiv:2312.13594*, 2023.

- [Laurençon *et al.*, 2023] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.
- [Lee *et al.*, 2020] Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. Contrastive learning with adversarial perturbations for conditional text generation. In *International Conference on Learning Representations*, 2020.
- [Li *et al.*, 2023a] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [Li *et al.*, 2023b] Chen Li, Yixiao Ge, Dian Li, and Ying Shan. Vision-language instruction tuning: A review and analysis. *arXiv preprint arXiv:2311.08172*, 2023.
- [Li *et al.*, 2023c] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [Li *et al.*, 2023d] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [Liang *et al.*, 2020] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3285–3292, 2020.
- [Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [Liu *et al.*, 2023a] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1, 2023.
- [Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.
- [Liu *et al.*, 2023c] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [Liu *et al.*, 2023d] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [OpenAI, 2023] R OpenAI. Gpt-4 technical report. *arXiv 2303.08774. View in Article*, 2:13, 2023.
- [Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [Ranzato *et al.*, 2015] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.
- [Suo *et al.*, 2023] Wei Suo, Mengyang Sun, Weisong Liu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. S3c: Semi-supervised vqa natural language explanation via self-critical learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2646–2656, 2023.
- [Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [Wang *et al.*, 2022] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khoshdel, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [Wang *et al.*, 2023a] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023.
- [Wang *et al.*, 2023b] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.
- [Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.