Learning Chain of Counterfactual Thought for Bias-Robust Vision-Language Reasoning

Yifeng Zhang, Ming Jiang, and Qi Zhao

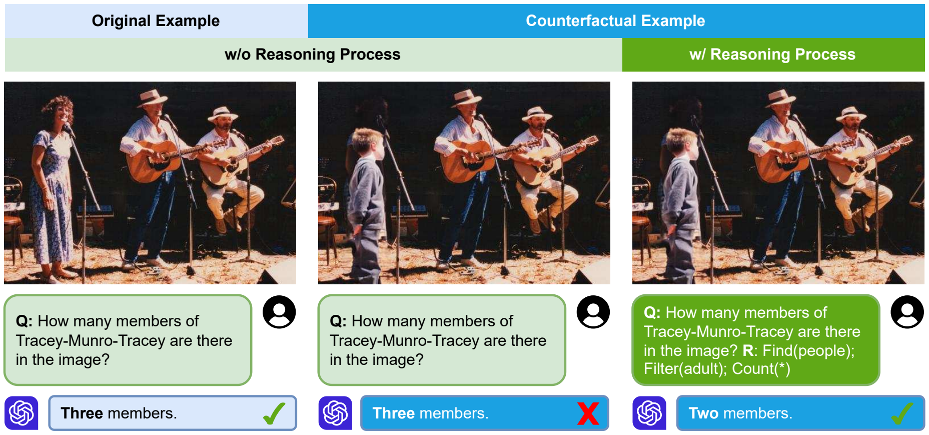
University of Minnesota, Minneapolis MN 55455, USA
{zhan6987, mjiang}@umn.edu, qzhao@cs.umn.edu

Abstract. Despite the remarkable success of large vision-language models (LVLMs) on various tasks, their susceptibility to knowledge bias inherited from training data hinders their ability to generalize to new scenarios and limits their real-world applicability. To address this challenge, we propose the Counterfactual Bias-Robust Reasoning (CoBRa) dataset that tackles knowledge bias by offering a novel collection of VQA examples designed to evaluate and mitigate bias in LVLMs. These examples encourage counterfactual thinking by providing edited knowledge graphs and image contents, with detailed annotations of reasoning processes to facilitate a comprehensive understanding of the examples. Based on the dataset, we introduce a Chain of Counterfactual Thought (CoCT) method that learns the bias-robust reasoning processes and provides in-context examples demonstrating how existing reasoning generalizes to counterfactual scenarios. This enables LVLMs to explicitly reason step-by-step rather than relying on biased knowledge, leading to more generalizable solutions. Our extensive evaluation demonstrates that CoCT outperforms existing approaches on tasks requiring reasoning under knowledge bias. Our work is available at https://github.com/SuperJohnZhang/CoBRa.

Keywords: Vision Language Model, Visual Reasoning, Counterfactual Thinking, Chain of Thought, Bias, Hallucination

1 Introduction

The recent emergence of large vision language models (LVLMs), exemplified by GPT-4V, highlights a transformative era in artificial intelligence by seamlessly integrating visual and linguistic reasoning. Despite their advancements, LVLMs are hindered by knowledge bias, challenging the realization of artificial general intelligence (AGI) [37]. Knowledge bias arises from inherent prejudices embedded within the training data. This bias can manifest in various forms, including social bias [5], cultural bias [48], linguistic bias [30], and factual bias [7], *etc.* Particularly, factual bias can critically influence the information LVLMs perceive and learn, causing the phenomenon of hallucination [7, 31], where LVLMs



Q: How many members of Tracey-Munro-Tracey are there in the image?

Three members. ✓

Q: How many members of Tracey-Munro-Tracey are there in the image?

Three members. ✗

Q: How many members of Tracey-Munro-Tracey are there in the image? **R**: Find(people); Filter(adult); Count(*)

Two members. ✓

Fig. 1: While GPT-4V accurately counts the Tracey-Munro-Tracy members in the example, its understanding and generalizability remain unclear. Counterfactual examples (*e.g.*, replacing a member with a child) reveal its struggle with new scenarios, suggesting bias-induced limitations. Our CoBRa dataset provides rich counterfactual examples and detailed reasoning annotations, enabling a novel CoCT method to mitigate bias in LVLMs and promote more generalizable multimodal understanding.

fabricate information or provide demonstrably false answers when encountering novel scenarios outside the scope of their training data.

Our work is driven by the aim of mitigating knowledge bias in LVLMs through counterfactual thinking [11]. This strategy involves crafting counterfactual examples, scenarios that deviate from the original situation in specific ways, necessitating the model to employ bias-robust step-by-step reasoning. As shown in Fig. 1, when prompted with a question like the number of members in Tracey-Munro-Tracey, without employing counterfactual thinking, the model (*e.g.*, GPT-4V) may rely solely on its prior knowledge, assuming all three individuals in the image are members of the trio. However, when confronted with a counterfactual scenario, the model can no longer provide the correct response unless guided by a bias-robust reasoning process. Our research delves into the data and methodological approaches that steer LVLMs towards more bias-robust reasoning processes, thereby reducing susceptibility to inherent knowledge bias.

In this paper, we introduce the Counterfactual Bias-Robust Reasoning (CoBRa) dataset, which presents two key novelties: (1) CoBRa features paired VQA examples with matched reasoning processes over varied knowledge contents (*i.e.*, three trio members *vs*. two members of the trio and a child, Fig. 1), sourced from potentially biased internet knowledge and counterfactual examples. Distinct from existing counterfactual data synthesis methods (*e.g.*, CSS [3], MUTANT [14]), the counterfactual examples are generated with systematically designed knowledge editing rules, editing both the knowledge graph and image contents. (2) CoBRa offers comprehensive annotations detailing the step-by-step reasoning processes, including the specific reasoning functions employed (*e.g.*, Find(), Filter(), Fig. 1) and relevant knowledge entities (*e.g.*, people, adult, Fig. 1). This rich annotation scheme allows researchers to not only evaluate the final answer of a model but also gain insights into its decision making, facilitating the identification of potential bias and promoting the development of more robust and explainable LVLMs.

Leveraging the counterfactual examples and annotated reasoning processes from CoBRa, we introduce Chain of Counterfactual Thought (CoCT), a novel chain-of-thought (CoT) method designed to mitigate the impact of knowledge bias on LVLMs. CoCT separates the reasoning functions from potentially biased factual knowledge, enabling the model to concentrate on bias-robust reasoning. By incorporating both original and counterfactual examples as in-context examples, LVLMs are prompted to maintain consistent reasoning across different scenarios. This approach significantly aids in mitigating bias influence, facilitating models' ability to generalize to novel situations.

In summary, the contributions of this paper are as follows:

1. Our CoBRa dataset introduces paired VQA examples with contrasting knowledge contents and comprehensive reasoning process annotations to facilitate the bias-robust reasoning of LVLMs.
2. Our CoCT method mitigates knowledge bias by separating the bias-robust reasoning functions from bias-sensitive knowledge and guiding consistent reasoning across original and counterfactual scenarios.
3. We demonstrate the effectiveness of CoBRa and CoCT in reducing biased responses through comprehensive experiments on multiple benchmarks. The results and findings offer valuable insights into combatting knowledge bias and hallucinations in LVLMs.

2 Related Work

2.1 Large Vision-Language Models

Our work builds upon the advancements in LVLMs, which have demonstrated significant advancements in artificial intelligence tasks, particularly due to their extensive pre-training on large-scale internet datasets for visual understanding [2]. Built on large language models (LLMs), their integration of the visual modality is often achieved through various architectural approaches: Models like Flamingo [1] combine a static vision component with a specialized LLM, while PaLM-E [10] directly integrates visual information into the powerful PaLM [6] architecture. Recent efforts focus on creating high-quality, diverse multimodal datasets from models like GPT-4 and GPT-4V. These datasets are then employed to fine-tune open-source LVLMs, such as LLaVA [32] and MiniGPT-4 [58], expanding their capabilities. Despite the excellent performance of LVLMs, they often tend to generate irrelevant or fabricated answers, namely hallucination hallucination [7,8,26,31]. Hallucinations arise from the model's over-reliance on bias within the training data, hindering generalization to new scenarios. In response, our CoBRa dataset provides paired VQA examples with matched reasoning processes upon different factual knowledge. CoCT leverages these counterfactual examples to guide LVLMs through bias-robust reasoning, offering a focused approach to address knowledge bias in LVLMs.

2.2 Bias Evaluation and Mitigation

Many comprehensive benchmarks have been proposed to evaluate a broad spectrum of cognitive abilities for LVLMs, such as recognition [12,33,49], OCR [36], language processing [34,56], and knowledge [35]. However, many of them often lack specific bias evaluation components. To combat bias, previous dataset efforts have focused on out-of-distribution VQAs [15,20] and hallucination-prone tasks [7,26,31]. They address specific aspects of bias but may not systematically evaluate all types of bias with sufficient examples. Counterfactual-based approaches [4,14] have been proposed to promote stronger connections between reasoning and inputs by generating synthetic images or questions. While they rely on simple masking or visual feature manipulation, our CoBRa dataset systematically generates diverse counterfactual examples by applying comprehensive knowledge editing rules based on knowledge graphs. CoBRa also provides step-by-step reasoning annotations for both original and counterfactual examples, guiding model training and empowering LVLMs in counterfactual thinking.

2.3 Chain-of-Thought Reasoning

Chain-of-thought (CoT) techniques have become a valuable tool for enhancing LLM reasoning abilities. They encourage LLMs to explicitly generate intermediate reasoning steps, leading to more accurate final results. CoT approaches fall into two main categories: zero-shot CoT [17,22,54,55], which relies on fine-tuned prompts without specific examples, and few-shot CoT [42,47], which provides multiple demonstrations of step-by-step reasoning for similar tasks. The former category advances in domain-specific performance due to the fine-tuning of large language models while the latter benefits flexibility in multi-domain deployment from the lightweight design. To improve CoTs in both categories, researchers are actively exploring ways by optimizing the reasoning process in prompts (*e.g.*, explicit decomposition of problems [21,57], employing calibration methods [13,27,46], *etc.*), or selecting high-quality demonstrations based on various metrics (*e.g.*, similarity [19,42], diversity [53], and complexity [13]). Recent efforts also generalize CoT to cross-modal CoT [34,54,55] for LVLMs. Unlike its precedents, our proposed CoCT is a novel cross-modal few-shot CoT method that optimizes both the reasoning process and the in-context demonstrations. It leverages counterfactual examples from the CoBRa dataset to predict bias-robust reasoning processes with separated logical steps (*i.e.*, functions) and knowledge entities (*i.e.*, parameters), and accumulates pairs of in-context examples with relevant reasoning processes over diverse factual knowledge, empowering LVLMs to achieve higher levels of bias-robustness in their decision making.

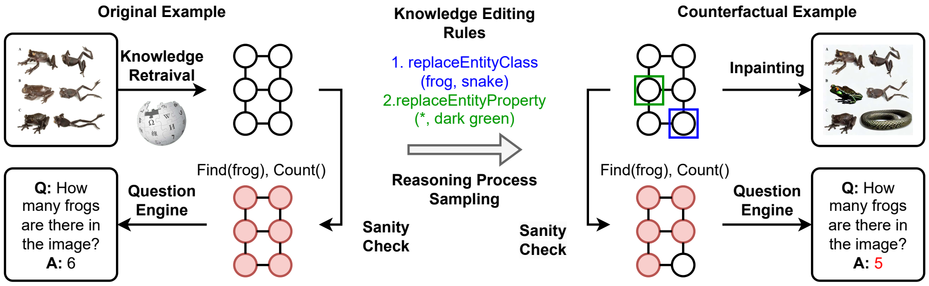

Original Example
Knowledge Retraival
Find(frog), Count()
Question Engine
Q: How many frogs are there in the image? A: 6
Sanity Check
Knowledge Editing Rules
1. replaceEntityClass (frog, snake)
2.replaceEntityProperty (*, dark green)
Reasoning Process Sampling
Sanity Check
Counterfactual Example
Inpainting
Find(frog), Count()
Question Engine
Q: How many frogs are there in the image? A: 5

Fig. 2: The two-stage pipeline of generating original and counterfactual examples with reasoning process annotations. **Left:** Build a knowledge graph from a Wikipedia image for original example generation and reasoning process sampling. **Right:** Applying comprehensive knowledge editing rules to generate a counterfactual knowledge graph, guiding the creation of counterfactual examples. This novel knowledge-graph-based approach forms the basis of CoBRa, enhancing LVLMs' reasoning and bias robustness.

3 Counterfactual Bias-Robust Reasoning Dataset

Our CoBRa dataset challenges LVLMs' reasoning abilities beyond knowledge bias by providing pairs of original and counterfactual examples and annotations of detailed reasoning processes. The construction of our CoBRa dataset follows a rigorous pipeline, involving data collection, knowledge retrieval and editing, reasoning process generation, question generation, and image generation. This pipeline comprises two stages (see Fig. 2): original example generation, and counterfactual example generation.

3.1 Generating Original Examples

The original examples are obtained from a collection of Wikipedia images and their corresponding textual descriptions [43]. Given an image, we employ state-of-the-art scene graph generation [16] and knowledge retrieval [51] methods to construct a comprehensive knowledge graph that organizes the relevant visual-linguistic information from the images and external sources (*e.g.*, attribution caption, sections of Wikipedia contents [43]). From the constructed knowledge graph, to generate the reasoning processes, along with the corresponding questions and answers, we first sample reasoning processes based on a pool of functions (*e.g.*, Find(), Count()) and match them with the parameters (*i.e.*, interconnected knowledge entities and relationships) sampled from the knowledge graph, arriving at parameterized reasoning processes (*e.g.*, Find(frogs), Count()). Specifically, reasoning functions denote the fundamental logical steps to arrive at a conclusion, while parameters are the specific knowledge contents employed within these functions. To examine the validity of the sampled reasoning process, we employ a series of sanity checkers following GQA-VD [52], such as checking the existence of reasoned knowledge entities, *etc*. Finally, upon successful

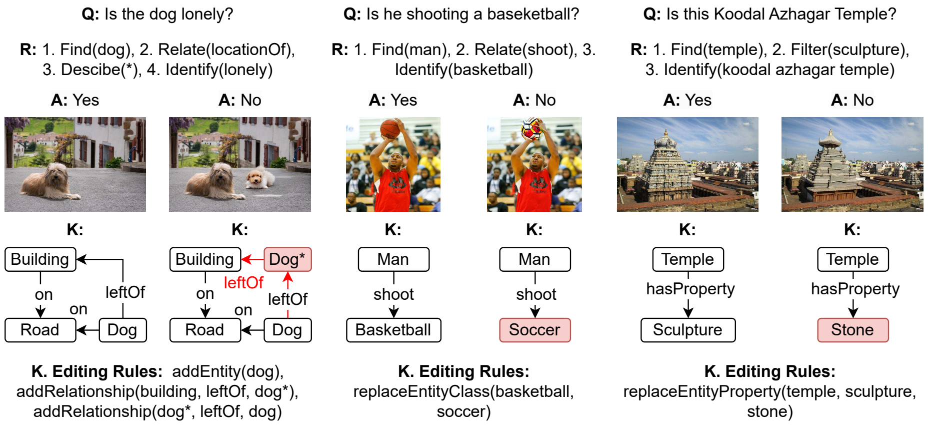

Q: Is the dog lonely?
R: 1. Find(dog), 2. Relate(locationOf), 3. Descibe(*), 4. Identify(lonely)
A: Yes
A: No
K:
Building
on
on
leftOf
Road
Dog
Building
Dog*
leftOf
on
on
leftOf
Road
Dog
K. Editing Rules: addEntity(dog), addRelationship(building, leftOf, dog*), addRelationship(dog*, leftOf, dog)
Q: Is he shooting a baseketball?
R: 1. Find(man), 2. Relate(shoot), 3. Identify(basketball)
A: Yes
A: No
K:
Man
shoot
Basketball
Man
shoot
Soccer
K. Editing Rules: replaceEntityClass(basketball, soccer)
Q: Is this Koodal Azhagar Temple?
R: 1. Find(temple), 2. Filter(sculpture), 3. Identify(koodal azhagar temple)
A: Yes
A: No
K:
Temple
hasProperty
Sculpture
Temple
hasProperty
Stone
K. Editing Rules: replaceEntityProperty(temple, sculpture, stone)

Fig. 3: To ensure the diversity of counterfactual examples, we leverage graph-based knowledge editing rules (*e.g.*, addEntity, addRelationship, replaceEntityClass, replaceEntityProperty *etc.*) Edited knowledge entities are highlighted in red.

validation, the reasoning processes and their associated facts are translated to questions and answers (*e.g.*, "Q: How many frogs are there in the image? A: 6") using the question engine of the GQA-VD dataset [52]. This process ensures that the questions are grounded in the visual-linguistic knowledge depicted in the knowledge graph.

3.2 Generating Counterfactual Examples

What distinguishes our method from previous studies is the generation of counterfactual examples. While existing counterfactual VQA methods simply apply masks or low-level manipulations to the image, we generate counterfactual examples by editing the knowledge graph and then inpainting the affected image regions, leading to more comprehensive and realistic results. As shown in Fig. 3, given an original example, we edit the relevant entities using multiple rules randomly sampled from a pool of comprehensive graph-based knowledge editing rules [40], such as replaceEntityClass(frog, snake) – to change the selected entity from a frog to a snake, or replaceEntityColor(dark green) – to change the color of the selected entity to dark green. The counterfactual images are generated from the edited knowledge graph using diffusion models [40, 41, 44]. To ensure the image quality, we apply a visual validator to examine whether the inpainted visual contents fit the desired scene and align with the edited knowledge graph. This is achieved by projecting the inpainted visual features and the updated knowledge graphs into the same latent space with CLIP [38] and comparing their similarity. Based on the updated knowledge graph, we then update the parameters of the original examples and apply the same sanity checkers/question engine, generating counterfactual reasoning processes, questions and answers.

Table 1: Statistics of CoBRa and related datasets. E. is short for edited images or QAs. Anns denotes the data annotation method. K. Anns and R. Anns denote the knowledge and reasoning annotations, respectively. T. denotes training set availability.

Dataset	# Imgs	# E. Imgs	# QAs	# E. QAs	Anns	K. Anns	R. Anns	T.
CoBRa	10k+	5k+	20k+	20k+	Auto	KG	Process	✓
VQA-CP v2 [15]	98k	0	220k	0	Manual	-	-	✓
GQA-OOD [20]	9.7k	0	53k	0	Auto	SG	Process	✓
HallusionBench [31]	1,129	504	1,129	1,129	Manual	-	-	-
Bingo [7]	308	140	370	370	Manual	-	-	-
MMMU [50]	11,550	0	11,550	0	Manual	-	-	-
ScienceQA [34]	10,332	0	21,208	0	Manual	-	Text	✓
MMBench [33]	2,974	0	2,974	0	Manual	-	-	-
MME [12]	1,187	0	1,457	1,457	Manual	-	-	-
MM-Vet [49]	200	0	218	155	Manual	-	-	-

This comprehensive approach ensures coherence in reasoning between the original and counterfactual examples despite factual knowledge shifts and bias.

3.3 Dataset Statistics

CoBRa introduces a novel approach to dataset creation for knowledge-based reasoning in counterfactual examples. It features 10,000 pairs of images, their corresponding question-answer sets, and reasoning annotations, meticulously edited using detailed knowledge graph annotations. With a fully automatic generation pipeline, this dataset can scale to more Wikipedia pages and currently maintains an additional training set with 44,604 pairs of examples, facilitating robust model training and testing across diverse scenarios. Tab. 1 highlights CoBRa's key benefits compared to existing datasets: (1) CoBRa stands out with its focus on comprehensive counterfactual scenarios, providing by far the only counterfactual image dataset with explicitly annotated knowledge graphs and reasoning processes. (2) To guarantee the diversity of knowledge for counterfactual examples, CoBRa leverages 254 compositional templates formed by a comprehensive set of graph-based knowledge editing rules. (3) CoBRa's scope extends beyond bias evaluation. With an automatic generation pipeline, CoBRa can generate more examples about a broader range of factual knowledge, allowing future methods to learn from our reasoning-grounded counterfactual examples.

As shown in Fig. 4, CoBRa promotes fair and generalizable AI models by incorporating a diverse range of topics and ensuring a balanced distribution of knowledge facts in the counterfactual examples. CoBRa's VQA examples span a diverse range of 12 natural and social science topics. The knowledge graph associated with each example has an average of 16 nodes and 46.9 edges. The dataset categorizes reasoning processes into five key skill sets: identification, comparison, classification, description, and arithmetic. Each reasoning process further decomposes into an average of 3.8 distinct reasoning functions (*e.g.*, Find(), Count()), enabling fine-grained analysis of a model's reasoning capabilities. This diversity

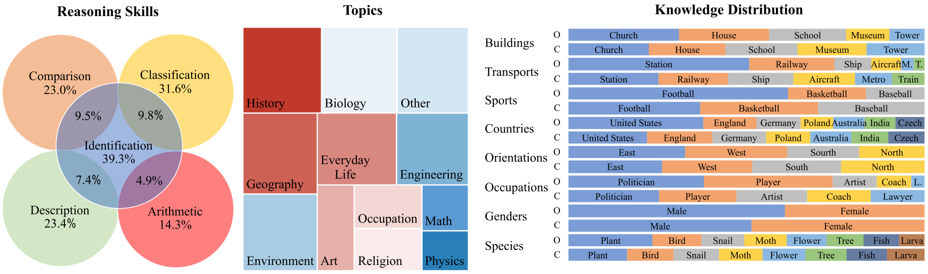


Fig. 4: Distributions of reasoning skills, topics, and knowledge examples. O stands for original examples and C stands for counterfactual examples.

of knowledge distribution offers a rich and balanced knowledge base for robust training and evaluation, leading to more reliable and inclusive models.

4 Chain of Counterfactual Thought

Chain of Thought (CoT) is a recent advancement in eliciting step-by-step reasoning behavior for LLMs and LVLMs. By exposing them to intermediate reasoning processes and high-quality in-context demonstrations, CoT not only generates more accurate answers but also provide an explanation. Inspired by its success, our proposed CoCT further advances with the counterfactual examples and reasoning annotations provided by the CoBRa dataset, empowering bias-robust reasoning for LVLMs. As shown in Fig. 5, it features two technical novelties: (1) Different from conventional CoT methods that treat logical steps and relevant knowledge as a whole (*e.g.*, rationales in DDCoT [55]), our CoCT explicitly separates them into reasoning functions and parameters. By leveraging the rich reasoning annotations from CoBRa, we train translation language models (TLM) [25], which excels at modeling relationships between linguistic sequences, and generate step-by-step bias-robust reasoning processes for LVLMs. (2) At the inference stage, CoCT retrieves contextually relevant examples from the training data based on the similarity of the reasoning process, with both the original and counterfactual examples being presented to the LVLM as in-context examples. This approach effectively guides the LVLM's counterfactual thinking to reason consistently across both scenarios.

4.1 Learning Bias-Robust Reasoning Processes

The proposed CoBRa dataset offers pairs of counterfactual examples along with their reasoning annotations, enabling the training of a dedicated model to predict the counterfactual "thought" for test examples. As shown in Fig. 5, CoCT adopts the TLM architecture [23], processing the LXMERT [45] embedding of the input to predict a sequence of reasoning steps $\mathcal{P} = \{\mathcal{P}^i\}$. Instead of predicting reasoning functions $\mathcal{P}_f$ and parameters $\mathcal{P}_p$ together, we employ two transformers in the

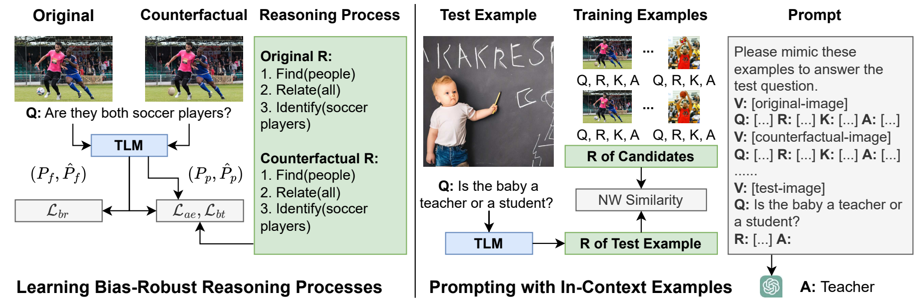

Original
Counterfactual
Reasoning Process
Q: Are they both soccer players?
TLM
$(P_f, \hat{P}_f)$
$(P_p, \hat{P}_p)$
$\mathcal{L}_{br}$
$\mathcal{L}_{ae}, \mathcal{L}_{bt}$
Original R:
1. Find(people)
2. Relate(all)
3. Identify(soccer players)
Counterfactual R:
1. Find(people)
2. Relate(all)
3. Identify(soccer players)
Learning Bias-Robust Reasoning Processes
Test Example
Training Examples
Prompt
Q, R, K, A
Q, R, K, A
Q, R, K, A
Q, R, K, A
R of Candidates
NW Similarity
Q: Is the baby a teacher or a student?
TLM
R of Test Example
Please mimic these examples to answer the test question.
V: [original-image]
Q: [...] R: [...] K: [...] A: [...]
V: [counterfactual-image]
Q: [...] R: [...] K: [...] A: [...]
......
V: [test-image]
Q: Is the baby a teacher or a student?
R: [...] A:
A: Teacher
Prompting with In-Context Examples

Fig. 5: Our CoCT method consists of two stages: **[Left]** Training a TLM to generate bias-robust reasoning processes, including functions and parameters. **[Right]** Prompting the LVLM with in-context examples retrieved from the training data, based on the similarity of reasoning processes.

TLM structure, each predicting one of them. The decoupling of functions and parameters allows models to flexibly adjust the parameters for counterfactual examples while preserving the functions, contributing to bias-robust reasoning.

The TLM training is supervised by ground-truth annotations of reasoning processes. To enhance generalizability and prevent overfitting, we simultaneously evaluate pairs of original and counterfactual examples: one question might have bias built-in, while the other is its bias-augmented counterparts. The TLM sees both questions at once, and learns to predict the correct reasoning steps for both.

To achieve this, the loss function is defined by extending the conventional denoising auto-encoding $\mathcal{L}_{ae}$ loss and back-translation loss $\mathcal{L}_{bt}$ of unsupervised neural machine translation system (NMT) [24, 25] with a bias-robust loss $\mathcal{L}_{br}$:

$$\mathcal{L} = \mathcal{L}_{br} + \mathcal{L}_{ae} + \mathcal{L}_{bt}. \tag{1}$$

The bias-robust loss encourages the original and counterfactual pair to maintain similar reasoning functions regardless of the parameters,

$$\mathcal{L}_{br} = -\sum_{i} \boldsymbol{p}_f^i \log(\hat{\boldsymbol{p}}_f^i), \tag{2}$$

where i is the positional index of the output and $\boldsymbol{p}_f^i$, $\hat{\boldsymbol{p}}_f^i$ are the predicted reasoning functions in a contrasting pair, respectively. The denoising auto-encoding and back-translation loss cooperate to ensure a robust translation. With a noise model that randomly drops tokens from the source (*i.e.*, questions) and the target domain (*i.e.*, functions and parameters), the denoising auto-encoding loss $\mathcal{L}_{ae}$ guarantees that the sequences at both domains can be reconstructed from their noisy versions, while the back-translation loss $\mathcal{L}_{bt}$ ensures that the noisy translation of sequences can be translated back into their original domain. With these losses, our models maintain stable performance in generating bias-robust reasoning processes.

Overall, this method addresses knowledge bias by translating the input question into a bias-robust reasoning process, minimizing its reliance on potentially biased factual knowledge. For detailed loss implementation, please refer to Supplementary Materials.

4.2 Prompting with In-Context Examples

For unbiased reasoning, as shown in Fig. 5 [Right], we search from CoBRa's unique original and counterfactual example pairs from the training dataset, and prompt LVLMs with those in-context examples.

Different from traditional in-context learning that relies on matching questions with similar examples from a large dataset, we consider the reasoning functions and parameters separately, where good in-context examples should maintain similarity in reasoning functions but diversity in parameters. With this strategy, CoCT does not rely on the knowledge similarity between in-context examples and the test example, hence allowing reasoning to be generalized to new scenarios.

Specifically, to measure the reasoning similarity between a test example's reasoning process and those in the CoBRa training dataset, we adopt the Needleman-Wunsch (NW) algorithm [28], a dynamic programming technique used to find the optimal alignment between two sequences by maximizing a similarity score while considering gaps. The original and counterfactual pairs with the highest similarity scores are then selected as the in-context examples, demonstrating how similar reasoning functions can be conducted to capture diverse visual contexts. Since the paired examples maintain diverse and balanced contexts in the original and counterfactual scenarios, they also help mitigate bias inherent in specific visual representations.

We integrate pairs of original and counterfactual examples into a structured prompt to guide bias-robust reasoning for LVLMs. As shown in Fig. 5, the prompt starts with a prompt head "*Please mimic these examples to answer the test question.*" to frame the VQA task, and then includes in-context examples following a specific structure: "*V: [original-image]; Q: [question]; R:[reason-process]; K:[knowledge]; A: [answer]*". Finally, the test example along with the predicted reasoning process is appended, leaving LVLMs to generate outputs. For detailed prompt examples, please refer to the Supplementary Materials.

5 Experiments

We comprehensively evaluate the effectiveness of our proposed CoBRa dataset and CoCT method in mitigating knowledge bias within LVLMs. We present experimental settings (Sec. 5.1) detailing the training and evaluation settings. We then conduct a quantitative evaluation (Sec. 5.2) by applying the CoCT to various benchmarks, comparing its performance against baselines, and measuring its impact on reducing bias. Finally, ablation studies (Sec. 5.3) systematically remove key components to analyze their contributions.

5.1 Experimental Settings

Datasets. To comprehensively evaluate the general reasoning and debiasing abilities of LVLMs, besides our proposed CoBRa dataset, we employed a diverse range of benchmark datasets encompassing various challenge categories: **Out-of-distribution Datasets:** GQA-OOD [20] and VQA-CPv2 [15] test models on scenarios where the answer distribution differs from training data, challenging them with uncommon or unexpected inputs. **Comprehensive Benchmarks:** MMBench [33] and ScienceQA with image contexts (SQA) [34], evaluating LVLMs across various domains and modalities, demanding reasoning across diverse contexts. **Hallucination Benchmarks:** Bingo (fact bias split) [7] and HallusionBench (HB) [31] contain questions requiring the model to imagine or hallucinate scenarios beyond the information provided in the image. Some subcategories intentionally shift the knowledge distribution to test LVLMs' capabilities to handle bias. This comprehensive selection ensures a rigorous evaluation of LVLMs' capabilities under various bias and reasoning challenges. For results on other LVLM benchmarks (*e.g.*, MME [12], MMMU [50], MM-Vet [49]) and hyperparameter tuning, please refer to the Supplementary Materials.
Compared Models. We evaluate our method against LVLM-based debiasing approaches, specifically focusing on two LVLMs: GPT-4V [2] and LLaVa-1.5 [32]. These models are tailored for vision-language tasks, leveraging pre-training on extensive image-text data to excel across various VLM benchmarks. We compare our CoCT method with three typical CoT methods: the clustering-based Auto-CoT [53], Active Prompting (AP) [9], and the rationale-generating DDCoT [55].
Training and Evaluation. Our method utilizes the CoBRa dataset to train a TLM and collect in-context examples for unbiased CoT reasoning. To prevent TLM from learning from seen examples, we split the dataset into training, validation, and test sets, ensuring evaluation validity and reliability. The trained TLM and reasoning-oriented heuristics are then assessed across all benchmarks.
Inference Time While CoT methods like AutoCoT, AP, and DDCoT improve reasoning, they potentially slow down the process by requiring large datasets or generating rationales. CoCT, focusing on core reasoning similarities, needs less data and achieves the fastest inference time (39ms) compared to others (AP: 41ms, AutoCoT: 48ms, DDCoT: 74ms), making it more suitable for real-world applications.

5.2 Quantitative Results

We quantitatively evaluate our method and current CoT approaches on LLaVA-1.5 and GPT-4V, presenting performance comparisons in Tab. 2. Our results reveal several findings:

- Adopting CoT methods (*e.g.*, AutoCoT, AP, DDCoT, and CoCT) leads to improved bias mitigation in LVLMs, particularly evident in significant enhancements over bias-evaluation benchmarks like CoBRa (16.1% on $\varDelta$), HB (9.3%), and Bingo (138.4%). This demonstrates the importance of incorporating step-by-step reasoning mechanisms to navigate biased shortcuts

Table 2: Comparison of model performance on multiple benchmarks. O and C refer to the original and counterfactual splits of CoBRa, respectively. Δ refers to the difference between two splits from CoBRa (*i.e.*, original and counterfactual) and GQA-OOD (*i.e.*, head and tail) respectively. All the experiments are conducted in a test set.

Method	CoBRa			VQA-CP	GQA-OOD		HB	Bingo	SQA	MMBench
	O	C	$\Delta\downarrow$		Tail	$\Delta\downarrow$				
LLaVA-1.5	63.1	18.9	44.2	50.4	47.8	9.9	29.0	12.5	66.8	64.3
+ AutoCoT [53]	63.3	19.6	43.7	53.2	48.2	8.8	29.4	18.8	67.9	65.1
+ AP [9]	63.4	21.4	42.0	52.8	48.3	8.9	30.1	18.8	68.4	65.2
+ DDCoT[1] [55]	74.4	23.9	50.5	55.8	48.7	9.1	30.8	23.4	**74.6**	68.4
+ CoCT (GQA)	73.5	33.4	40.1	**56.2**	**50.1**	8.4	31.4	23.4	70.8	67.9
+ CoCT (CoBRa)	**76.4**	**38.3**	**38.1**	56.0	49.2	**8.4**	**32.0**	**29.8**	71.3	**68.9**
GPT-4V	78.2	36.7	41.5	54.8	49.3	9.4	37.3	9.4	71.3	68.9
+ AutoCoT [53]	79.1	37.9	41.2	55.7	49.7	9.1	37.8	10.9	73.2	77.8
+ AP [9]	78.8	38.0	40.8	56.3	49.5	9.4	37.4	12.5	74.1	78.2
+ DDCoT [55]	79.1	38.4	40.7	59.8	50.2	8.8	37.7	23.4	**80.2**	78.3
+ CoCT (GQA)	78.9	40.1	38.8	58.8	**50.9**	8.9	38.5	26.6	74.2	68.3
+ CoCT (CoBRa)	**79.3**	**41.8**	**37.5**	**60.4**	50.6	**8.5**	**41.2**	**31.4**	75.8	**78.5**

for more equitable and robust reasoning. More importantly, high-performing CoT approaches also outperform bare LVLMs on comprehensive benchmarks (*e.g.*, +6.7% on SQA, +7.1% on MMBench), underscoring the crucial role of bias mitigation in bolstering LVLMs' performance and robustness across diverse evaluation settings.

– Our CoCT outperforms existing CoT approaches (*i.e.*, AutoCoT, AP, DDCot) across all bias-evaluation benchmarks (*i.e.*, CoBRa, VQA-CP, GQA-OOD, HB, Bingo) and MMBench, emphasizing its effective usage of the bias-robust reasoning process. While the rationale generation method DDCoT achieves higher performance on SQA, its fine-tuning for the rationale generation on a large language model [39] exacerbates the bias, leading to limited performance on bias-evaluation datasets. Differently, CoCT provides a lightweight solution by explicitly decoupling reasoning functions from potentially biased knowledge with TLMs, achieving more accurate reasoning and better generalizability.

– Our CoBRa dataset, providing counterfactual examples with diverse knowledge but similar reasoning functions, significantly enhances adaptability and generalization across various reasoning tasks. The combination of CoCT and CoBRa examples outperforms CoCT with GQA examples, achieving the top performance on 8/10 and 9/10 metrics with LLaVA-1.5 and GPT-4V, respectively. It is noteworthy that leveraging GQA examples is advantageous on GQA-OOD and VQA-CP due to the same source (*i.e.*, MS COCO [29], GQA [18]) for images and questions of GQA examples.

[1] Different from few-shot CoT methods, DDCoT fine-tunes a large language model on the CoT annotations of SQA to generate reasoning rationales.

Table 3: Ablation study of TLM and bias-robust Loss in CoCT on multiple datasets.

Method	CoBRa-O	CoBRa-C	$\Delta\downarrow$	HB	Bingo	SQA
Baseline	63.1	18.9	44.2	29.0	12.5	66.8
w/o TLM	69.8	28.5	41.3	31.5	18.8	68.9
w/o Bias-Robust Loss $\mathcal{L}_{br}$	74.5	34.2	40.3	31.8	26.6	70.4
Full	**76.4**	**38.3**	**38.1**	**32.0**	**29.8**	**71.3**

Table 4: Impacts of different settings of in-context examples. K.Sim stands for examples selected by the similarity of factual knowledge.

Setting	CoBRa-O	CoBRa-C	$\Delta\downarrow$	HB	Bingo	SQA
Zero-Shot	68.5	27.1	41.4	29.4	18.8	67.9
Random	72.3	31.1	41.2	30.7	23.4	70.1
K.Sim	72.7	31.2	41.5	30.5	18.8	70.8
Original	72.9	32.1	40.8	31.1	23.4	69.3
Counterfactual	73.5	33.7	39.8	31.4	23.4	69.8
Both	**76.4**	**38.3**	**38.1**	**32.0**	**29.8**	**71.3**

5.3 Ablation Study

We explore the impact of various components through ablation studies on the LLaVA-1.5 model, including the TLM that predicts the reasoning process (Tab. 3), the choices of in-context examples (Tab. 4) and the design of prompts (Tab. 5).

TLM and Bias-Robust Loss. The TLM model of CoCT is trained on the CoBRa examples with the bias-robust loss $\mathcal{L}_{br}$, explicitly generating step-by-step reasoning processes for the training and test examples. As shown in Tab. 3, ablating this model effectively removes the model's ability to perform CoT reasoning, leading to a drastic performance drop among all four benchmarks (*i.e.*, CoBRa $\Delta\downarrow$: +3.2; HB: -0.5, Bingo: -11.0; SQA: -3.6). This result suggests the crucial role of the step-by-step reasoning process in eliciting the reasoning capabilities of LVLMs across different scenarios. Besides, incorporating a TLM without the bias-robust loss in CoCT not only degrades overall performances, but also contributes to a steeper performance drop on counterfactual examples (*i.e.*, CoBRa-C (-4.1) *vs.* CoBRa-O (-1.9)), indicating that the bias-robust loss is able to mitigate the influence of potential bias embedded within the training data and generalize LVLM more effectively to counterfactual scenarios.

In-Context Examples. The choice of in-context examples is also highly relevant to the model's reasoning capabilities. To evaluate how in-context examples impact the reasoning of LVLMs, as shown in Tab. 4, we evaluate CoCT under six different settings: (1) `Zero-Shot`: only a template example to regulate the output format of CoT; (2) `Random`: randomly selected examples; (3) `K.Sim`: examples with high knowledge semantic similarity; (4-6) `Original`/`Counterfactual`/`Both`: only original/counterfactual or both examples with high reasoning similarity computed by NW [28]. Our results reveal four observations:

Firstly, integrating in-context examples (2-6) yields improved performance over the zero-shot setting (1) across various benchmarks (CoBRa-O, CoBRa-C,

Table 5: Impacts of different prompt components.

Prompt	CoBRa-O	CoBRa-C	$\Delta\downarrow$	HB	Bingo	SQA
w/o Reasoning Processes	68.5	24.3	44.2	29.4	14.1	67.8
w/o Knowledge Graphs	71.4	29.9	41.5	31.4	26.6	69.2
Full	**76.4**	**38.3**	**38.1**	**32.0**	**29.8**	**71.3**

HB, SQA), demonstrating their ability to harness LVLM reasoning potential efficiently. Secondly, employing different similarity metrics (3-6) enables LVLMs to access more relevant contexts compared to random examples (2), resulting in enhanced reasoning performance in specific scenarios; for instance, K.Sim achieves superior performance (SQA: 70.8) compared to the random setting (SQA: 70.1). Thirdly, while K.Sim excels in benchmarks requiring diverse knowledge (SQA: 70.8), our reasoning similarity examples (4-6) demonstrate greater effectiveness in counterfactual scenarios (CoBRa-C: >32.1), highlighting the significance of separating reasoning from biased knowledge for bias-robust decision-making. Lastly, incorporating both original and counterfactual examples (6) significantly improves performance over single-type examples (4, 5) across all bias-evaluation benchmarks (CoBRa $\Delta\downarrow$: -2.7, HB: +0.9, Bingo: +6.4), emphasizing the crucial role of contrastive comparison in bias mitigation.

Prompt Design. Tab. 5 show the impact of different prompt components on LVLM's reasoning capabilities, including reasoning processes and knowledge graphs. Overall, combining both components yields the highest performance (*i.e.*, 38.1 on CoBRa, 32.0 on HB, *etc.*), highlighting their effectiveness. Besides, omitting reasoning processes causes the most significant performance decline (*e.g.*, 38.1 → 44.2 on Δ for CoBRa), indicating their crucial role in guiding LVLMs for bias-robust reasoning and improved generalizability.

6 Conclusion

This paper tackled the pervasive issue of knowledge bias in LVLMs. It offers two key contributions for achieving bias-robust reasoning: CoBRa is a novel dataset featuring counterfactual examples and detailed reasoning annotations. CoBRa facilitates studying LVLMs' reasoning and empowers the development of bias-mitigating techniques. Powered by CoBRa's counterfactual examples and reasoning processes, CoCT is a novel approach for mitigating knowledge bias through step-by-step reasoning. CoCT achieves this by decoupling the core reasoning functions from potentially biased knowledge and prompting LVLMs with counterfactual examples that share the reasoning steps but differ in factual details. Their effectiveness, demonstrated through comprehensive experiments, paves the way for next-generation LVLMs that prioritize fairness and reliability. Future work may explore CoCT's application to broader vision-language tasks while integrating explainability techniques for deeper insights into its reasoning process and further combatting bias in AI systems.

Acknowledgment: This work is supported by NSF Grants 2143197 and 2227450.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
4. Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10800–10809 (2020)
5. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3043–3054 (2023)
6. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: Palm: Scaling language modeling with pathways (2022)
7. Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., Yao, H.: Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. arXiv preprint arXiv:2311.03287 (2023)
8. Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J.: Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495 (2023)
9. Diao, S., Wang, P., Lin, Y., Zhang, T.: Active prompting with chain-of-thought for large language models (2023)
10. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model (2023)
11. Facione, P.A., et al.: Critical thinking: What it is and why it counts. Insight assessment **1**(1), 1–23 (2011)
12. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)

13. Fu, Y., Peng, H., Sabharwal, A., Clark, P., Khot, T.: Complexity-based prompting for multi-step reasoning (2023)
14. Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: Mutant: A training paradigm for out-of-distribution generalization in visual question answering. arXiv preprint arXiv:2009.08566 (2020)
15. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017)
16. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1969–1978 (2019)
17. Guo, Y., Nie, L., Wong, Y., Liu, Y., Cheng, Z., Kankanhalli, M.: A unified end-to-end retriever-reader framework for knowledge-based vqa. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2061–2069 (2022)
18. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
19. Jie, Z., Lu, W.: Leveraging training data in few-shot prompting for numerical reasoning. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 10518–10526. Association for Computational Linguistics, Toronto, Canada (Jul 2023). `https://doi.org/10.18653/v1/2023.findings-acl.668`, `https://aclanthology.org/2023.findings-acl.668`
20. Kervadec, C., Antipov, G., Baccouche, M., Wolf, C.: Roses are red, violets are blue... but should vqa expect them to? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2776–2785 (2021)
21. Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., Sabharwal, A.: Decomposed prompting: A modular approach for solving complex tasks (2023)
22. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners (2023)
23. Lample, G., Conneau, A.: Cross-lingual language model pretraining. CoRR **abs/1901.07291** (2019), `http://arxiv.org/abs/1901.07291`
24. Lample, G., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. CoRR **abs/1711.00043** (2017), `http://arxiv.org/abs/1711.00043`
25. Lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M.: Phrase-based & neural unsupervised machine translation. CoRR **abs/1804.07755** (2018), `http://arxiv.org/abs/1804.07755`
26. Li, J., Cheng, X., Zhao, W.X., Nie, J.Y., Wen, J.R.: Halueval: A large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6449–6464 (2023)
27. Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.G., Chen, W.: Making large language models better reasoners with step-aware verifier (2023)
28. Likic, V.: The needleman-wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne pp. 1–46 (2008)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–

ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)

30. Liu, F., Bugliarello, E., Ponti, E.M., Reddy, S., Collier, N., Elliott, D.: Visually grounded reasoning across languages and cultures. arXiv preprint arXiv:2109.13238 (2021)
31. Liu, F., Guan, T., Li, Z., Chen, L., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multimodality models. arXiv preprint arXiv:2310.14566 (2023)
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
33. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
34. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS) (2022)
35. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
36. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR (2019)
37. Qu, Y., Wei, C., Du, P., Che, W., Zhang, C., Ouyang, W., Bian, Y., Xu, F., Hu, B., Du, K., et al.: Integration of cognitive tasks into artificial general intelligence test for large models. arXiv preprint arXiv:2402.02547 (2024)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
39. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. CoRR **abs/1910.10683** (2019), http://arxiv.org/abs/1910.10683
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
42. Rubin, O., Herzig, J., Berant, J.: Learning to retrieve prompts for in-context learning (2022)
43. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2443–2449. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463257, https://doi.org/10.1145/3404835.3463257
44. Stan, G.B.M., Wofk, D., Fox, S., Redden, A., Saxton, W., Yu, J., Aflalo, E., Tseng, S.Y., Nonato, F., Muller, M., et al.: Ldm3d: Latent diffusion model for 3d. arXiv preprint arXiv:2305.10853 (2023)
45. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)

46. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations (2022)
47. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023)
48. Yin, D., Li, L.H., Hu, Z., Peng, N., Chang, K.W.: Broaden the vision: Geo-diverse visual commonsense reasoning. arXiv preprint arXiv:2109.06860 (2021)
49. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
50. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)
51. Zhang, Y., Jiang, M., Zhao, Q.: Explicit knowledge incorporation for visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1356–1365 (2021)
52. Zhang, Y., Jiang, M., Zhao, Q.: New datasets and models for contextual reasoning in visual dialog. In: European Conference on Computer Vision. pp. 434–451. Springer (2022)
53. Zhang, Z., Zhang, A., Li, M., Smola, A.: Automatic chain of thought prompting in large language models (2022)
54. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models (2023)
55. Zheng, G., Yang, B., Tang, J., Zhou, H.Y., Yang, S.: Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models (2023)
56. Zheng, K., Chen, X., Jenkins, O.C., Wang, X.E.: Vlmbench: A compositional benchmark for vision-and-language manipulation (2022)
57. Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., Chi, E.: Least-to-most prompting enables complex reasoning in large language models (2023)
58. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023)