Understanding Visual Detail Hallucinations of Large Vision-Language Models

Xiaoxi Sun[1], **Jianxin Liang**[1], **Yueqian Wang**[1], **Huishuai Zhang**[1*], **Dongyan Zhao**[1,2*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]State Key Laboratory of General Artificial Intelligence
sunxiaoxi@stu.pku.edu.cn,{liangjx, wangyueqian, zhanghuishuai, zhaody}@pku.edu.cn

Abstract

Understanding small visual objects is crucial in fields such as video surveillance, remote sensing, and autonomous driving. In this paper, we investigate the capability of advanced large vision-language models (LVLMs) to recognize and interpret small objects in visual data. To this end, we curate a specialized dataset for evaluating fine-grained visual hallucinations, incorporating two object categories and three types of hallucinations. First, we assess 11 state-of-the-art LVLMs, yielding several key insights, as anticipated, LVLMs perform significantly worse on queries related to small objects compared to regular-sized ones, with performance on regular objects proving to be an unreliable predictor of that on small objects. This finding underscores the need for dedicated research on fine-grained visual hallucinations. Second, we evaluate three training-free methods: Scaffold, Chain of Thought (CoT), and Image Resizing, all of which result in varying degrees of improvement. Furthermore, we conduct a series of detailed ablation studies on the visual encoders of Eagle-X5, examining their performance across fine-grained visual hallucination tasks. Our findings reveal that ConvNeXt architecture is critical for object existence recognition tasks. In contrast, for mitigating other types of hallucinations, integrating information from multiple visual encoders is significantly more effective than relying on a single encoder. These results highlight several promising directions for advancing small object recognition with LVLMs.

1 Introduction

In recent years, large language models (LLMs) [Achiam *et al.*, 2023; Touvron *et al.*, 2023] achieve significant breakthroughs in natural language processing, excelling in tasks such as language understanding [Hendrycks *et al.*, 2020] and generation [Fernandes *et al.*, 2023]. These advancements also propel the emergence of large vision-language models (LVLMs) [Liu *et al.*, 2024a; Achiam *et al.*, 2023; Caffagni *et al.*, 2024], which further enhance performance on multimodal tasks, e.g., image captioning [Agrawal *et al.*, 2019] and visual question answering [Goyal *et al.*, 2017].

However, hallucinations in LVLMs, where generated outputs deviate from the input images, pose significant challenges to their deployment in critical domains such as healthcare and autonomous driving [Li *et al.*, 2023b; Bai *et al.*, 2024]. Addressing these hallucinations is essential for ensuring the reliability of multimodal models. First and foremost, establishing a reliable evaluation framework is the top priority. Existing benchmarks, such as CHAIR [Rohrbach *et al.*, 2018], an early framework for assessing object hallucinations in image captioning tasks, and POPE [Li *et al.*, 2023b] that frames hallucination evaluation as a binary classification task, primarily assess general object understanding. However, these benchmarks inadequately capture LVLMs' capabilities in scenarios involving small objects, which are critical for applications like video surveillance [Zhu *et al.*, 2021] and autonomous driving [Pang *et al.*, 2020]. Our preliminary experiments in Section 4 reveal that low hallucination rates for regular objects do not necessarily correlate with low hallucination rates for small objects, underscoring a critical gap in current evaluation methodologies.

To address this gap, we curate a specific dataset for evaluating and analyzing visual detail hallucinations. The dataset categorizes scenes into two types: detailed and regular, and includes three key hallucination types: existence, color, and position. This dataset enables a comprehensive analysis of visual detail hallucinations and supports meaningful comparisons between detailed and regular scenes.

Using this dataset, we perform extensive experiments on 11 state-of-the-art models, yielding several notable observations. First, the use of a mixture of visual encoders proves to be highly effective, consistently outperforming single-encoder approaches. Second, increasing the resolution of input images significantly enhances small-object recognition for models capable of processing images at varying resolutions. Third, the Chain of Thought (CoT) prompting method substantially improves performance on complex visual tasks, such as determining positional relationships. Additionally, we evaluate another training-free method, Scaffold [Lei *et al.*, 2024], which adds anchor positions to input images. The results show mixed performance across the three hallucination types.

[*] Corresponding authors: Huishuai Zhang and Dongyan Zhao.

To gain deeper insights, we investigate the role of individual visual encoders within the mixture used by the EAGLE-X5 [Shi *et al.*, 2024] model, focusing on their impact on small-object recognition. Our findings reveal that the ConvNeXt encoder predominantly determines performance for existence hallucination questions. For other hallucination types, the model effectively integrates features from multiple visual encoders, leading to consistent improvements compared to using a single encoder. These results highlight the complementary strengths of different encoders in addressing various hallucination challenges.

In summary, our key contributions are as follows:

- **A Specialized Dataset**: We curate a dataset for visual detail hallucination assessment, encompassing two scene types and three hallucination categories, enabling robust evaluation.
- **Model and Method Analysis**: We evaluate multiple models and methods, revealing that most models struggle with detailed visual information. We show that a combination of visual experts and larger resolutions mitigate visual detail hallucinations.
- **Analytical Insights**: We investigate other design factors influencing visual detail hallucinations, including the contributions of various visual encoders in mixed architectures, the impact of image resizing, and the role of training data distribution, offering valuable guidance for future research.

This work establishes a pipeline for curating a dataset specifically to evaluate visual detail hallucinations in large vision-language models (LVLMs). We assess the performance of current LVLMs on this dataset and systematically evaluate several improvement strategies, providing in-depth analyses of their effectiveness. Additionally, we investigate underlying factors contributing to visual detail hallucination issues, offering valuable insights and identifying key directions for future research and efforts to enhance LVLM performance on small-object recognitions.

2 Related Work

This work focuses on evaluating the visual detail hallucination of large vision-language models. For clarity, we introduce the related work with the following three categories.

2.1 Large Vision-Language Models

Large Vision-Language Models(LVLMs) typically comprise three key components: an LLM backbone for user interaction, one or more visual encoders, and vision-to-language adapter modules [Caffagni *et al.*, 2024]. The visual encoder is critical for perceiving visual information. Early LVLMs [Li *et al.*, 2023a; Liu *et al.*, 2024b] commonly use CLIP [Radford *et al.*, 2021] as the visual encoder. For example, LLaVA [Liu *et al.*, 2024b] projects CLIP-encoded visual features into the text space through linear projection or MLP, integrating them with textual embeddings for further processing. Subsequent research has explored LVLMs with a hybrid visual encoder structure[Shi *et al.*, 2024; Lin *et al.*, 2023; Fan *et al.*, 2024], demonstrating that combining multiple visual encoders significantly improves the model's visual capabilities. However, these structures are typically limited to processing images at fixed resolutions. Recently, some studies [Wang *et al.*, 2024; Wu *et al.*, 2024] have enabled models to handle arbitrary resolutions. For instance, Qwen2-VL [Wang *et al.*, 2024] enhances Visual Transformer(ViT) [Dosovitskiy, 2020] by replacing absolute position embeddings with 2D-RoPE, enabling the model to effectively capture the two-dimensional positional information of images. Overall, models with different structures exhibit varying capabilities in perceiving visual details. The effectiveness of methods for mitigating visual detail hallucinations depends on the model's structure. In this work, we conduct detailed experiments and analyses to explore these differences.

2.2 Hallucination Benchmarks for LVLMs

In vision-language models, hallucinations [Bai *et al.*, 2024] refer to scenes where the generated output does not align with the input image. Existing research primarily focuses on object hallucinations, and so do the evaluation benchmarks. CHAIR [Rohrbach *et al.*, 2018], an early work predating the advent of large models, evaluates object hallucinations in image captioning tasks. It evaluates how many generated words correspond to objects present in the image by leveraging ground truth sentences and object segmentations within a fixed set of objects. However, CHAIR's applicability is confined to traditional image captioning tasks with predefined object sets, making it unsuitable for assessing the diverse outputs of contemporary vision-language models. More recently, POPE [Li *et al.*, 2023b] has emerged as a widely used benchmark for evaluating object hallucinations. POPE transforms hallucination evaluation into a binary classification task by asking yes-or-no questions about relevant objects. While POPE is reliable and flexible for evaluating existence hallucinations, it narrowly focuses on this specific type of hallucination, limiting its broader applicability. Additionally, several other benchmarks [Lovenia *et al.*, 2023; Liu *et al.*, 2023a] have been introduced to provide a more structured evaluation of hallucinations in multimodal models. In comparison, we focus on visual detail hallucinations and propose a pipeline for constructing related datasets and evaluation methods.

3 Evaluation Dataset Construction

In this section, we introduce our curated dataset, which evaluates hallucinations across three categories, further divided into small and regular objects. The small object dataset evaluates the model's ability to handle visual detail hallucinations, while the regular object dataset assesses hallucinations in standard scenarios. By comparing performance across both datasets, we quantify the model's performance degradation when handling small objects. We begin by defining the problem and categorizing hallucinations. Next, we describe the methods for data collection and filtering.

3.1 Task Definition and Hallucination Categories

We define ***Small Objects*** based on the small object detection task [Cheng *et al.*, 2023] in computer vision to identify vi-

Model	Scope	Existence				Color				Position				Total	PP
		Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score		
LLaVA-1.5-7B	$\mathcal{S}$	73.7	74.6	71.8	73.2	59.0	56.5	78.0	65.5	55.8	53.5	88.0	66.5	188.5	35.3
	$\mathcal{R}$	86.0	81.8	92.7	86.9	71.0	65.7	88.0	75.2	66.8	60.6	96.0	74.3	223.8	
LLaVA-1.5-13B	$\mathcal{S}$	74.3	77.2	68.9	72.8	62.0	60.0	72.0	65.5	57.3	54.2	**94.5**	**68.9**	193.6	35.3
	$\mathcal{R}$	86.6	83.1	92.0	87.3	75.0	71.9	82.0	76.6	67.3	61.0	96.0	74.6	228.9	
Deepseek-VL-7B-Chat	$\mathcal{S}$	75.3	81.6	65.3	72.6	63.0	65.1	56.0	60.2	52.3	51.9	63.0	56.9	190.6	44.1
	$\mathcal{R}$	88.4	88.2	88.7	88.4	83.0	83.7	82.0	82.8	63.3	59.5	83.0	69.3	234.7	
Qwen-VL-Chat-7B	$\mathcal{S}$	73.6	74.8	71.3	73.0	68.0	64.5	**80.0**	71.4	58.5	55.1	92.0	**68.9**	200.1	32.9
	$\mathcal{R}$	86.7	82.6	92.7	87.4	83.0	83.7	82.0	82.8	63.3	59.5	83.0	69.3	233.0	
CogVLM2-19B	$\mathcal{S}$	65.7	**94.0**	33.5	49.4	64.0	81.8	36.0	50.0	61.8	69.7	41.5	52.0	191.5	50.6
	$\mathcal{R}$	90.1	95.5	84.3	89.5	85.0	92.7	76.0	83.5	67.0	67.7	65.0	66.3	242.1	
SPHINX	$\mathcal{S}$	73.3	79.6	62.6	70.1	65.0	63.2	72.0	67.3	62.8	60.2	75.5	67.0	201.1	41.9
	$\mathcal{R}$	88.0	87.5	88.7	88.1	81.0	77.2	88.0	82.2	74.0	68.3	89.5	77.5	243.0	
Eagle-X5-13B-Chat	$\mathcal{S}$	**81.0**	79.0	**84.3**	**81.6**	70.0	67.9	76.0	**71.7**	63.5	60.0	81.0	**68.9**	**214.5**	**30.0**
	$\mathcal{R}$	89.5	84.7	96.3	90.1	82.0	77.6	90.0	83.3	73.0	66.9	91.0	77.1	244.5	
Deepseek-VL2-Tiny	$\mathcal{S}$	76.5	73.4	83.1	77.9	63.0	72.4	42.0	53.2	63.8	73.1	43.5	54.5	203.3	46.5
	$\mathcal{R}$	86.0	79.5	97.1	87.4	89.0	95.3	82.0	88.2	74.8	77.0	70.5	73.6	249.8	
Qwen2-VL-Chat-7B	$\mathcal{S}$	72.3	91.7	49.1	63.9	**73.0**	**82.9**	58.0	68.2	**64.5**	**74.6**	44.0	55.3	209.8	35.3
	$\mathcal{R}$	90.1	92.0	87.8	89.9	82.0	86.4	76.0	80.9	73.0	77.7	64.5	70.5	245.1	
Qwen-VL-MAX	$\mathcal{S}$	67.3	95.3	36.5	52.7	65.0	72.7	48.0	57.8	60.0	85.7	24.0	37.5	192.3	43.9
	$\mathcal{R}$	88.7	95.5	81.2	87.8	80.0	81.3	78.0	79.6	67.5	84.3	43.0	57.0	236.2	
GPT-4o	$\mathcal{S}$	75.1	78.3	69.3	73.6	63.0	72.4	42.0	53.2	60.0	62.5	50.0	55.6	198.1	37.9
	$\mathcal{R}$	89.2	88.7	89.8	89.2	85.0	88.9	80.0	84.2	61.8	62.2	60.0	61.1	236.0	

Table 1: **Accuracy(%), Recall(%), Precision(%), F1(%) of different models on three categories.** $\mathcal{S}$ represents small objects, which correspond to visual detail hallucinations, while $\mathcal{R}$ represents regular objects. **Total** is calculated as the sum of accuracies across the three types of hallucinations, while **Performance Penalty (PP)** represents the difference in Total scores between the two scenarios. The best scores for small objects are highlighted in **bold**, and the best scores for regular objects are indicated with underlining.

sual details. Specifically, objects with a bounding box area smaller than 1024 pixels or occupying less than 2% of the total image area are classified as *Small* ($\mathcal{S}$), while all others are categorized as *Regular* ($\mathcal{R}$). The classification of visual detail hallucinations aligns with that of regular object hallucinations. Since POPE [Li *et al.*, 2023b] only evaluates existence hallucinations and does not assess other categories, we expand the categorization of object hallucinations into three types:

- **Existence**: Misperceptions about the existence of an object, such as describing an object that does not exist.
- **Color**: Misperceptions about the primary color of an object.
- **Position**: Misperceptions about the relative positioning of two objects.

3.2 Data Collection

Data Annotation Inspired by the previous work [Li *et al.*, 2023b; Rohrbach *et al.*, 2018], we choose MSCOCO [Lin *et al.*, 2014] as the source for our dataset. Since the dataset lacks annotations for color and positional relationships, we generate this information separately. For the color category, we manually annotate the data by using bounding boxes to locate selected objects and label their primary colors. For positional relationships, we derive the information from the bounding boxes of the objects, specifically by analyzing the relative positions of their center points. The positional relationships are categorized as: *["bottom right," "top right," "bottom left," "top left"]*.

Small Object Data Filtering Our evaluation dataset is divided into two parts based on object size: small objects and regular objects. Additionally, it is categorized into three types of hallucinations. For each type, the number of data points is balanced between the two object categories, with the distinction based on whether filtering is applied. Different filtering methods are employed for each type of hallucination:

- **Existence**: All detected bounding boxes for each object are recorded. If the largest bounding box area falls below a given threshold, the image is classified as containing small objects.
- **Color**: An object is classified as a small object only if it has a **unique** bounding box and its area is below the specified threshold.
- **Position**: All **unique** objects in an image are paired, and at least one pair must include a small object for the combination to be considered valid.

Each small object must have a **unique standalone bounding box** (*Color* category), and only **non-duplicate** objects in the image will be paired (*Position* category). The validation set of MS COCO2014 contains fewer than 20 images with a total pixel count below 51,200 (1,024 pixels occupy 2% of the total 51,200 pixels). A bounding box area smaller than 1,024 pixels typically occupies less than 2% of the image area, reducing computational costs during filtering. We thus adopt the bounding box area as the filtering criterion. Through this filtering process, we obtain an annotated dataset of small object images and generate an unfiltered dataset of the same size for regular objects. Using these datasets, we create yes-or-no questions from the existing images to evaluate hallucinations.

Question Generation We employ a template-based approach to generate questions for each image by incorporating the corresponding annotated information: object name, color, and the positional relationship between two objects.

Model	Method	Scope	Existence				Color				Position			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
LLaVA-1.5-7B	Baseline	$\mathcal{S}$	**73.7**	74.6	**71.8**	**73.2**	**59.0**	**56.5**	78.0	65.5	55.8	53.5	88.0	**66.5**
		$\mathcal{R}$	86.0	81.8	92.7	86.9	71.0	65.7	88.0	75.2	66.8	60.6	96.0	74.3
	Scaffold	$\mathcal{S}$	72.4	76.9	64.0	69.9	55.0	53.3	80.0	64.0	**56.3**	**54.0**	85.0	66.0
		$\mathcal{R}$	86.7	84.5	89.8	87.1	72.0	65.7	92.0	76.7	62.3	57.7	91.5	70.8
	CoT	$\mathcal{S}$	-	-	-	-	**59.0**	55.4	**92.0**	**69.2**	53.3	52.3	73.0	61.0
		$\mathcal{R}$	-	-	-	-	66.0	59.8	98.0	74.2	59.3	56.5	80.0	66.3
	Image Resizing	$\mathcal{S}$	72.2	**78.8**	60.7	68.6	57.0	54.1	**92.0**	68.1	49.5	49.7	**95.0**	65.3
		$\mathcal{R}$	86.9	85.3	89.1	87.1	66.0	60.5	92.0	73.0	51.8	50.9	99.5	67.3
Qwen2-VL-Chat-7B	Baseline	$\mathcal{S}$	72.3	91.7	49.1	63.9	73.0	82.9	58.0	68.2	64.5	74.6	44.0	55.3
		$\mathcal{R}$	90.1	92.0	87.8	89.9	82.0	86.4	76.0	80.9	73.0	77.7	64.5	70.5
	Scaffold	$\mathcal{S}$	71.5	**93.2**	46.5	62.1	69.0	**88.0**	44.0	58.7	62.8	**83.1**	32.0	46.2
		$\mathcal{R}$	90.5	95.1	85.3	90.0	81.0	91.9	68.0	78.2	70.5	82.0	52.5	64.0
	CoT	$\mathcal{S}$	-	-	-	-	73.0	78.0	64.0	70.3	65.5	71.8	51.0	59.6
		$\mathcal{R}$	-	-	-	-	84.0	85.4	82.0	83.7	75.0	75.8	73.5	74.6
	Image Resizing	$\mathcal{S}$	**75.1**	91.0	**55.6**	**69.0**	75.0	72.7	**80.0**	**76.2**	**70.0**	**75.0**	**60.0**	**66.7**
		$\mathcal{R}$	91.1	92.1	89.9	91.0	83.0	85.1	80.0	82.5	75.3	75.4	75.0	75.2
Eagle-X5-13B-Chat	Baseline	$\mathcal{S}$	81.0	79.0	**84.3**	**81.6**	**70.0**	67.9	76.0	71.7	63.5	60.0	**81.0**	68.9
		$\mathcal{R}$	89.5	84.7	96.3	90.1	82.0	77.6	90.0	83.3	73.0	66.9	91.0	77.1
	Scaffold	$\mathcal{S}$	77.3	80.5	72.1	76.0	**70.0**	67.2	**78.0**	**72.2**	60.5	58.5	72.0	64.6
		$\mathcal{R}$	89.1	86.6	92.4	89.4	83.0	77.0	94.0	84.7	71.5	66.9	85.0	74.9
	CoT	$\mathcal{S}$	-	-	-	-	68.0	**71.4**	60.0	65.2	60.5	57.7	78.5	66.5
		$\mathcal{R}$	-	-	-	-	80.0	78.8	82.0	80.4	68.8	62.7	92.5	74.7
	Image Resizing	$\mathcal{S}$	**81.3**	**81.3**	81.3	81.3	**70.0**	67.2	**78.0**	**72.2**	**65.0**	**61.4**	**81.0**	**69.8**
		$\mathcal{R}$	90.3	86.3	95.7	90.8	82.0	75.8	94.0	83.9	72.3	66.8	88.5	76.1

Table 2: **Experiment results of different methods on three kinds of models.** The **Scaffold** method adds anchor positions to input images as visual prompts to improve model performance. The **CoT** method enables the model to generate bounding boxes as intermediate information for better reasoning. The **Image Resizing** method upscales images by a factor of 2 before inputting them into the model.

For positive examples, we directly use the correct information. To generate negative examples, we follow a method similar to POPE's [Li *et al.*, 2023b] negative example generation. First, we collect all object names (or colors, or positional relationships) from the positive examples. Then, for each instance, we randomly or adversarially select an object name (or color, or positional relationship) from this collection that differs from the instance to create a negative example. The templates we used are as follows:

```
Existence: Is there {article} {object} in
the image?
Color: Is the main color of the {object} in
the image {color}?
Position: Is {object1} to the {direction}
of {object0} in the image?
```

We create the evaluation dataset using this generation method. The statistical details for the data are summarized in Table 3. The data examples can be found in Appendix A.3.

Metrics We evaluate the model using four metrics: Accuracy, Precision, Recall, and F1 Score. Additionally, we calculate the total accuracy across the three hallucination categories, with separate scores for regular and small objects. The difference between these two scores highlights the performance degradation when processing visual details.

4 Evaluation of Models and Methods

In this section, we first evaluate the performance of 11 state-of-the-art LVLMs on the annotated dataset, leading to several notable observations. Next, we examine the effectiveness of three training-free data augmentation methods in mitigating hallucinations.

Category	# Questions	# P / # N
Existence	3000	1500/1500
Color	100	50/50
Position	400	200/200

Table 3: **The statistical summary of the dataset.** "# Questions" denotes the number of questions in the corresponding category, while "# P/ # N" indicates the number of positive and negative examples for the respective category.

4.1 Evaluation of LVLMs on Visual Detail Hallucination

We evaluate several LVLMs, including LLaVA-1.5 [Liu *et al.*, 2024a], DeepSeek-VL [Lu *et al.*, 2024; Wu *et al.*, 2024], Qwen-VL [Bai *et al.*, 2023; Wang *et al.*, 2024], CogVLM2 [Hong *et al.*, 2024], SPHINX [Lin *et al.*, 2023], Eagle-X5 [Shi *et al.*, 2024], and GPT-4o [Achiam *et al.*, 2023]. These models exhibit diverse architectural designs. Specifically, LLaVA, DeepSeek-VL, Qwen-VL, and CogVLM2 employ CLIP or its variants as visual encoders, which map images to a fixed resolution for processing. Sphinx and Eagle utilize a mixture of visual encoders with varying architectures. Deepseek-VL2 and Qwen2-VL adopt distinct approaches to enable models to process images of any resolution. Qwen-VL-MAX and GPT-4o are closed-source commercial models. We assess these models on our dataset to evaluate their performance in answering questions about visual detail hallucination. The results are shown in Table 1.

Our initial observations reveal that all models, including GPT-4o, experience varying degrees of performance degradation when transitioning from regular object scenes to those with smaller objects. Moreover, a model's ability to perceive

Methods	Scope	Accuracy	
		Color	Position
Qwen-VL-MAX	$\mathcal{S}$	62.0	60.0
	$\mathcal{R}$	81.0	67.5
w. CoT	$\mathcal{S}$	59.0($\downarrow$ *3.0*)	73.5 ($\uparrow$ *13.5*)
	$\mathcal{R}$	83.0 ($\uparrow$ *2.0*)	85.3($\uparrow$ *17.8*)

Table 4: **Experiment results of CoT method on Qwen-VL-MAX.** CoT significantly improves the accuracy of answering positional questions but shows minimal or negative impact on color questions.

small objects does not consistently align with its performance on normal scenes. For example, CogVLM excels in recognizing regular objects, outperforming Eagle in the existence and color categories. However, its performance on small objects declines significantly, particularly in the existence category, where it underperforms relative to other models. This discrepancy suggests that improving a model's performance on regular objects does not necessarily lead to enhanced performance on visual details, highlighting the importance of developing datasets specifically designed to evaluate visual detail-related hallucinations. Furthermore, the performance of the same model varies across different categories, underscoring the necessity of categorical evaluation.

Among these models, Sphinx and Eagle, which employ a mixture of visual encoders, demonstrate strong performance on both small and regular objects. Eagle achieves the highest total accuracy score on the small object evaluation set. This success is largely attributed to the fact that most models typically rely on a single ViT-based visual encoder. While this approach excels at aggregating long-range interactions due to its training process, it often struggles to capture neighboring dependencies [Lin *et al.*, 2023]. The mixed visual encoder partially addresses this limitation. We will provide further analysis about it in Section 5.1. Additionally, Deepseek-VL2 and Qwen2-VL outperform models that can only handle fixed resolutions, as they reduce information loss when high-resolution images are mapped to a fixed resolution.

4.2 Evaluation of Training-Free Methods for Hallucination Mitigation

We evaluate several training-free methods across different models. **Scaffold** [Lei *et al.*, 2024] is a visual prompt method that overlays a dot matrix within the image to serve as visual information anchors, utilizing multidimensional coordinates as textual positional references. **Chain of Thought(CoT)** [Wei *et al.*, 2023] typically prompts models to generate the reasoning process before outputting the final answer. In this task, we specifically prompt models to utilize bounding boxes of objects mentioned in the question as intermediate information for generating the reasoning process. For the color category, the model first detects the bounding box of the object and then determines whether the color within the bounding box is correct. For the position category, the model detects the bounding boxes of two objects, calculates their center points, and judges the positional relationship based on the relative positions of these points. For the existence category, detecting bounding boxes is more challenging than directly determining existence, so we did not conduct experiments in this category. The prompts we used can be found in Appendix A.1. **Image Resizing** is a straightforward approach to addressing small object hallucinations. We use bicubic interpolation to resize the images by a factor of 2 before evaluation. We conduct experiments on three methods using three distinct models: LLaVA, which employs a standard visual encoder; Qwen2-VL, which handles images of any resolution; and Eagle, which integrates multiple visual experts. The experimental results are summarized in Table 2.

The Scaffold method enhances performance in specific models and hallucination categories compared to the baseline, particularly in regular object scenes. However, its effectiveness is limited for small objects, often resulting in performance degradation. This decline occurs because adding visual prompts directly to an image has little impact on regular objects but can obscure small objects, making them harder for the model to perceive.

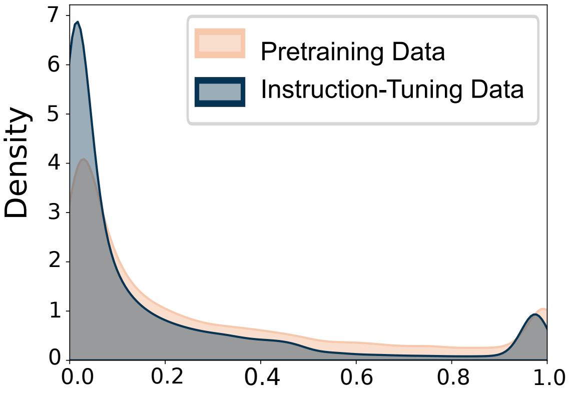


Figure 1: **Object Area Distribution of LLaVA's Training Data.** The **x-axis** represents the ratio of the object's bounding box area to the total image area, while the **y-axis** represents the corresponding density. In both data categories, small objects exhibit a high probability density.

For Qwen2-VL, which handles images of any resolution, upsampling provides additional visual details, achieving the best performance across all three hallucination categories. The image resizing method also enhances Eagle's accuracy in answering existence and positional relationship hallucination questions. This improvement stems from Eagle's mix of visual experts, which includes models like SAM [Kirillov *et al.*, 2023] that can process higher-resolution images. In contrast, the LLaVA model exhibits suboptimal performance with the image resizing method. This is primarily due to the visual encoder's preprocessing step, which resizes images to a fixed resolution. Upsampling the image before this resizing does not change the final resolution but degrade image quality, resulting in a lower-quality representation compared to the original.

The improvement from the CoT method is more limited across all models, partly because smaller models have less advanced language capabilities and struggle to comprehend complex instructions. To explore this, we test the CoT method on Qwen-VL-MAX, a commercial model with

Combination	Scope	Accuracy		
		Existence	Color	Position
A	$\mathcal{S}$	79.1	64.0	59.3
	$\mathcal{R}$	87.9	75.0	63.5
AB	$\mathcal{S}$	79.6	66.0	60.3
	$\mathcal{R}$	86.8	76.0	64.5
ABC	$\mathcal{S}$	79.6	66.0	60.5
	$\mathcal{R}$	86.7	76.0	64.5
ABCD	$\mathcal{S}$	80.8	68.0	61.8
	$\mathcal{R}$	89.5	81.0	73.3
ABCDE	$\mathcal{S}$	81.0	70.0	63.5
	$\mathcal{R}$	89.5	82.0	73.0

Table 5: **Accuracy of EAGLE with different combinations of visual encoding features.** In this table, visual encoders are denoted as follows: A for ConvNeXT, B for CLIP, C for SAM, D for EVA-02-L, and E for Pix2Struct. By incorporating information from different encoders, the model's performance improves to varying extents across the respective categories, thereby highlighting the contribution of each encoder in addressing specific categories of problems.

stronger language capabilities. The results in Table 4 reveal that the CoT method significantly improves performance in answering positional relationship questions on Qwen-VL-MAX. However, it provides only minor enhancements for color-related questions and even exhibits a slight performance decline when addressing small object issues. We analyze the outputs and derive the following insights. First, Qwen-VL-MAX performs poorly in object existence perception, with over 40% of responses in the color category stating '*can't detect the object*.' This significantly degrades performance on small objects. Furthermore, for data points where the model with CoT answers "*can't detect the object*", it answers "*no*" 95% of the time without CoT, indicating that object detection capability directly impacts color perception. Additionally, we find that CoT with bounding box as intermediate information highly relies on the models' ability to perceive object existence. Overall, the CoT method does not directly enhance the model's visual perception capabilities. Second, the primary advantage of CoT lies in its ability to infer information beyond direct visual perception. For example, we observe a case where the model struggles to detect the bounding box of a very small phone. However, by identifying the position of the person holding the phone, it infers the likely location of the phone and correctly answers the question. This example can be found in Appendix A.2. In summary, the image resizing method mitigates hallucinations by providing the model with more detailed visual information, whereas the CoT method leverages the model's reasoning ability to infer additional information for hallucination mitigation.

5 Analysis of Design Factors

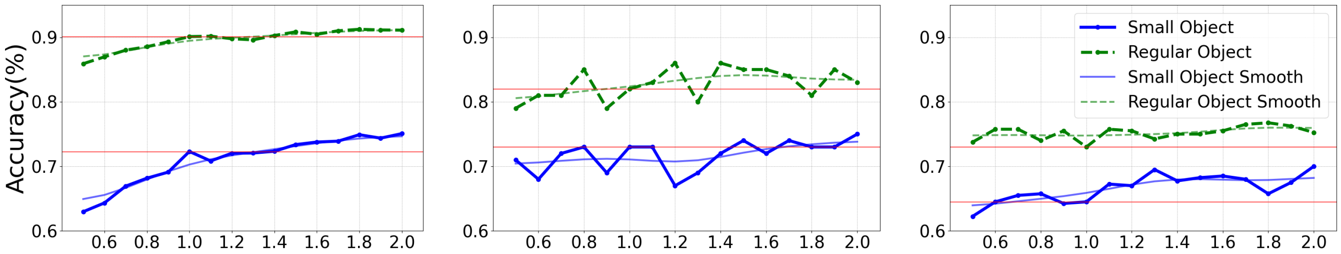
In this section, we analyze the architecture of mixed visual encoders and examine the distribution of object area across different training stages using LLaVA-1.5 as an example, to infer their potential effect on hallucinations related to small objects. Then, we analyze how the resize factor affects models that can handle images of any resolution.

5.1 Influence of Mixed Visual Encoders

In Section 4, we observe that models with mixed visual experts generate fewer hallucinations. To further explore this phenomenon, we conduct training-free experiments using the Eagle-X5-13B-Chat model, which exhibits superior performance in Table 1. This model processes visual information through a mixture of vision encoders combined via channel concatenation and utilizes a three-stage training recipe including vision expert pre-alignment. Firstly, we evaluate Eagle's performance using features from a single visual encoder across all three hallucination categories. Secondly, we combine information from multiple encoders and analyze their individual contributions to solving specific problems. To assess the impact of each encoder, we selectively mask its encoded features, disabling its contribution both individually and in combination. We test five visual encoders: CLIP, ConvNeXt [Woo *et al.*, 2023], SAM [Kirillov *et al.*, 2023], Pix2Struct [Lee *et al.*, 2023], and EVA-02-L [Fang *et al.*, 2024]. While ConvNeXt is based on a convolutional network, the others utilize the ViT architecture and are trained with distinct processes or for specific tasks.

From Table 6, we observe that task-specific models like SAM (segmentation) and Pix2Struct (OCR) perform poorly when relying solely on features from a single visual encoder for existence hallucination problems. While CLIP's encoding outperforms these two, it still underperforms compared to models trained exclusively with CLIP, such as LLaVA-v1.5. In contrast, ConvNeXt's encoding yields the best results for existence hallucinations, even surpassing some models listed in Table 1. Moreover, augmenting ConvNeXt's encoding with other encoders' information does not substantially improve performance. This suggests that the model relies mainly on ConvNeXt's information and treats other encoder data as auxiliary. This also explains why performance drops more significantly when using only CLIP's encoding, compared to models trained with CLIP, as the model treats CLIP's information as secondary. For other types of hallucination problems, ConvNeXt's encoding still provides an advantage, though the effect is less pronounced compared to existence hallucinations. Incorporating information from encoders such as CLIP and EVA-02-L results in significant improvements, indicating their substantial contribution to solving these problems.

The results reveal several interesting insights. When multiple visual encoders process an image, each encoder generates distinct encoded features due to variations in architecture, training procedures, and objectives. These encoders can be likened to an LVLM equipped with multiple "eyes", akin to human vision, where each eye may perceive the same scene differently. Similar to the human brain, which processes visual information asymmetrically from both eyes [Khan and Crawford, 2001], a visual language model exhibits a "dominant" encoder after joint training. In such cases, the model prioritizes information from one encoder, analogous to the dominant eye, while treating input from the other encoders as supplementary. This structure shows potential for reducing small object hallucinations, suggesting that targeted training could enable the model to leverage different visual inputs for small object detection while using others for general scenes.



(a) Existence

(b) Color

(c) Position

Figure 2: **Influence of resizing factors on the accuracy of responses to three categories of hallucination problems.** The **x-axis** represents the resize factor. The two red lines in each figure represent the performance with original resolution in both $\mathcal{R}$(regular) and $\mathcal{S}$(small). For *Existence* and *Position*, accuracy improves as the image size increases, while no significant correlation is observed for *Color* category.

Visual Expert	Scope	Accuracy		
		Existence	Color	Position
ConvNeXt	$\mathcal{S}$	79.1	64.0	59.3
	$\mathcal{R}$	87.9	75.0	63.5
CLIP	$\mathcal{S}$	57.8	62.0	50.0
	$\mathcal{R}$	68.6	66.0	50.5
SAM	$\mathcal{S}$	50.0	64.0	52.8
	$\mathcal{R}$	49.4	70.0	51.8
EVA-02-L	$\mathcal{S}$	63.5	57.0	52.8
	$\mathcal{R}$	67.2	63.0	57.5
Pix2Struct	$\mathcal{S}$	48.5	62.0	51.3
	$\mathcal{R}$	46.4	69.0	49.5

Table 6: **Accuracy of EAGLE in answering three categories of hallucination using only a single visual encoding feature.** Notably, ConvNeXt excels in handling the existence hallucination problem, making it the primary contributor to the response for this issue. For other types of hallucinations, multiple visual encoders contribute to varying extents.

5.2 Small Object Distribution in Training Data

We analyze the distribution of small objects in its training data. The LLaVA-1.5 training process involves two main stages: pre-training for modality alignment and visual instruction tuning. The pre-training utilizes a 558K subset of the LAION-CC-SBU dataset with BLIP captions, while the instruction tuning phase incorporates from COCO2017, GQA [Hudson and Manning, 2019], OCR-VQA [Mishra *et al.*, 2019], TextVQA [Sidorov *et al.*, 2020], and VisualGenome [Krishna *et al.*, 2017]. Since OCR-VQA and TextVQA focus on text within images, we exclude them from our analysis.

We extract the objects from the training data using part-of-speech tagging methods and apply Grounding DINO [Liu *et al.*, 2023b] to obtain their bounding boxes. We then analyze the distribution of bounding box areas for all mentioned objects. The results in Figure 1 indicate that the probability density of small objects is relatively high in both pre-training and instruction tuning data. Therefore, the proportion of small objects in the training data does not appear to be the primary cause of the model's tendency to hallucinate small objects.

5.3 Influence of Resizing Factors

Furthermore, we explore the influence of different resize factors on accuracy. Specifically, we set 16 factors ranging from 0.5 to 2.0 incremented by 0.1. We resize the image with a factor r: when $r > 1$, the image is upsampled with bicubic interpolation, and when $r < 1$, the image is downsampled with Lanczos interpolation. While LLaVA-1.5 preprocesses images by resizing them to a fixed resolution, we explore the use of Qwen2-VL here, which can handle images of any resolution, to provide a more flexible analysis.

The experimental results are summarized in Figure 2. For *Existence* category (Figure 2a), a clear positive correlation between the resize factor and accuracy is observed, particularly for small objects. As the image size increases, accuracy consistently improves, indicating that upscaling enhances the model's ability to detect small objects. In contrast, for regular objects, accuracy remains stable, suggesting that the resize factor has little effect on their detection. For *Color* category (Figure 2b), no significant correlation between resize factor and accuracy is observed for either small or regular objects. The performance fluctuates without a clear trend, implying that color recognition is less sensitive to changes in image size. For *Position* category (Figure 2c), accuracy improves as the resize factor increases, particularly for small objects. This trend highlights the importance of upscaling in enhancing spatial reasoning for small objects, whereas the performance for regular objects remains consistent.

In summary, the resize factor significantly affects the existence and position categories, particularly for small objects, but has minimal impact on the color category.

6 Conclusion

In this work, we curate a dataset to evaluate and analyze visual detail hallucinations, filtering data based on automatic or manual annotations for corresponding categories. We evaluate multiple models and methods, revealing that architectures like the mixture of visual experts and visual encoders capable of handling images of any resolution effectively mitigate these hallucinations. Additionally, the CoT and Image Resizing methods exhibit strong potential in mitigating visual detail hallucinations. Through an ablation study on model architecture, we find that ConvNeXt architecture is critical for object existence recognition tasks, and integrating information from multiple visual encoders is significantly more effective than relying on a single encoder for mitigating other types of hallucinations.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Agrawal *et al.*, 2019] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[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.

[Bai *et al.*, 2024] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[Caffagni *et al.*, 2024] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*, 2024.

[Cheng *et al.*, 2023] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fan *et al.*, 2024] Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, Ming Zhang, Caishuang Huang, Rui Zheng, Zhiheng Xi, Yuhao Zhou, Shihan Dou, Junjie Ye, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Mousi: Poly-visual-expert vision-language models, 2024.

[Fang *et al.*, 2024] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

[Fernandes *et al.*, 2023] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668, 2023.

[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.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[Hong *et al.*, 2024] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[Hudson and Manning, 2019] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019.

[Khan and Crawford, 2001] Aarlenne Z Khan and J Douglas Crawford. Ocular dominance reverses as a function of horizontal gaze angle. *Vision research*, 41(14):1743–1748, 2001.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[Lee *et al.*, 2023] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[Lei *et al.*, 2024] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2023b] 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.

[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.

[Lin *et al.*, 2023] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

[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. In *The Twelfth International Conference on Learning Representations*, 2023.

[Liu *et al.*, 2023b] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[Lovenia *et al.*, 2023] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*, 2023.

[Lu *et al.*, 2024] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

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

[Pang *et al.*, 2020] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing*, 30:207–219, 2020.

[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.

[Rohrbach *et al.*, 2018] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[Shi *et al.*, 2024] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.

[Sidorov *et al.*, 2020] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[Wei *et al.*, 2023] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[Woo *et al.*, 2023] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023.

[Wu *et al.*, 2024] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

[Zhu *et al.*, 2021] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021.