FiVL: A Framework for Improved Vision-Language Alignment through the Lens of Training, Evaluation and Explainability

Anonymous ACL submission

Abstract

Large Vision Language Models (LVLMs) have achieved significant progress in integrating visual and textual inputs for multimodal reasoning. However, a recurring challenge is ensuring these models utilize visual information as effectively as linguistic content when both modalities are necessary to formulate an accurate answer. We hypothesize that hallucinations arise due to the lack of effective visual grounding in current LVLMs. Furthermore, current vision-language benchmarks are not specifically measuring the degree to which the answer require the visual input. This limitation makes it challenging to confirm that the image is truly necessary, particularly in tasks like visual question answering. In this work, we introduce FiVL, a novel method for constructing datasets designed to train LVLMs for enhanced visual grounding and also evaluate their effectiveness in achieving it. We demonstrate the value of our datasets through three approaches. First, we introduce a novel training task based on our augmented training dataset, resulting in better performance than the baseline. Second, we present benchmarks to assess the model's ability to use image as substantive evidence, rather than relying solely on linguistic priors. Finally, we identify attention heads with the strongest vision-language alignment, enabling explainability on visual-driven hallucinations. The dataset and code will be publicly available.

1 Introduction

Recent advancements in large language models have led to the integration of non-linguistic information through multimodal perception and generation, culminating in the development of Large Vision Language Models (LVLM). These models effectively bridge visual comprehension and linguistic reasoning, offering a unified approach to multimodal understanding and instruction-following (Liu et al., 2024c; Peng et al., 2023; Wang et al., 2024; OpenAI, 2024). However, despite their apparent adeptness in visual perception, LVLMs still face the challenge of "hallucination" — where the model generates semantically plausible yet factually incorrect information that is inconsistent with the image. This issue possibly arises from the imbalance of visual data compared to text data during training, which limits the model's ability to overcome preconceptions inherited from the underlying LLM (Liu et al., 2024a). A common approach to mitigate this issue has been to introduce a visual grounding mechanism to the model (Chen et al., 2023; Peng et al., 2023; Wang et al., 2024; Rasheed et al., 2024; Zhang et al., 2024b,a).

Visual grounding aims to achieve more precise alignment between visual attention and semantic concepts within the model. A common method for grounding involves using bounding boxes, represented as a sequence of numerical numbers, to specify a particular region of an image. This enables the user to query specific parts of the image and the model to reference image locations within its generated response (Chen et al., 2023; Peng et al., 2023; Wang et al., 2024). Bounding boxes, however, are coarse coordinates and unable to highlight objects or abstract concepts in finer detail. Recent work have addressed these concerns by applying pixel-level grounding through the use of segmentation masks instead (Rasheed et al., 2024; Zhang et al., 2024b,a).

Training models with pixel-level grounding requires datasets that provide fine-grained visual alignment between images and text. However, such datasets are scarce, and prior work have often constructed custom datasets alongside model development (Zhang et al., 2024a; Rasheed et al., 2024; Ma et al., 2024; Zhang et al., 2024b). Additionally, prior works use grounding datasets only for training and overlook the importance of alignment datasets for evaluation. To address these challenges, we introduce FiVL, a novel Framework towards

Improved Vision-Language Alignment, for constructing datasets with visual-concept alignment. We demonstrate the usefulness of these datasets through a novel training approach as well as a method that evaluates and interprets the visual-language alignment capability in LVLMs.

Our main contributions are as follows:

- We introduce FiVL, a framework designed to augment multimodal datasets with vision-alignment capabilities. Through comprehensive human and automated evaluations of the datasets produced by FiVL, we demonstrate their reliability.
- Using the FiVL training dataset, we introduce a novel training task that jointly trains text and vision tokens. Leveraging this task, we fine-tuned an LVLM model that outperforms the baseline across several downstream tasks.
- Through FiVL evaluation datasets, we use a perturbation-based approach to assess the vision-alignment capability of LVLMs and introduce visual reliance score. This score shows a strong correlation with overall model performance, going beyond a specific subset of benchmarks.
- We leverage our framework in order to gain more insights into the internal mechanisms of LVLMs by identifying attention heads with the strongest vision-language alignment capabilities, as demonstrated in (Aflalo et al., 2022). This approach enables the exploration of vision-based hallucinations.

2 Related Work

LVLM and Visual Grounding. Building upon LLMs, LVLMs extend their capabilities to a multimodal context by incorporating visual perception into the generation process, with notable models such as GPT-4o (OpenAI, 2024), LLaVA (Liu et al., 2024c), Qwen2-VL (Wang et al., 2024), and many others (Dai et al., 2023; Zhu et al., 2024; Chen et al., 2024), demonstrating advanced visual reasoning ability. Additionally, some LVLMs employ grounding mechanisms to enhance multimodal interaction by allowing the model to reference specific regions of an image. This visual grounding has been achieved though the prediction of bounding boxes coordinates, as seen in models such as Kosmos-2 (Peng et al., 2023), Shikra (Chen et al., 2023), BuboGPT (Zhao et al., 2023), Ferret (You et al., 2024), Qwen2-VL (Wang et al., 2024), and Groma (Ma et al., 2024). To obtain a fine-grained localization of objects and semantic concepts pixel-level grounding has subsequently proposed in models such as Llava-Grounding (Zhang et al., 2024a), GLaMM (Rasheed et al., 2024), and GROUNDHOG (Zhang et al., 2024b). Unlike other grounding methods which learn to treat bounding box coordinates as part of the "language" of the model like Kosmos-2 and Shikra, or GROUNDHOG that was trained to output mask along with generated text, FiVL's training process explicitly ties image tokens to their vocabulary text representation, creating fine-grained alignment that enhances visual grounding beyond simple region prediction, while preserving a text-based interface. The resulting model not only improves accuracy but can also be utilized to produce segmentation maps.

Visually Grounded Datasets. Training LVLMs require large-scale visual instruction-following data (Liu et al., 2024c). However, these datasets focus on the task of visual and language reasoning and generally do not have fine-grained image segmentation annotations. Prior work have mainly constructed custom datasets to train their respective grounded LVLM models. In (Ma et al., 2024), a custom dataset, Groma Instruct, was constructed by prompting GPT-4V to generate grounded conversations based on 30K samples with region annotations from COCO (Lin et al., 2014) and VG (Krishna et al., 2017). Llava-Grounding (Zhang et al., 2024a) curated the Grounded Visual Chat (GVC) dataset by matching class labels of ground truth bounding boxes from COCO to noun phrases in conversations from LLaVA-Instruct-150K (Liu et al., 2024c) using GPT-4. The Grounding-anything Dataset (GranD) was specifically constructed to train GLaMM (Rasheed et al., 2024) and utilized an object detection model to obtain visual entities that were then used to generate grounded dense captions through an LLM. A grounded visual instruction tuning dataset, M3G2, was proposed to train the GROUNDHOG model (Zhang et al., 2024b). There, the authors curated a dataset consisting of 2.5M text-image pairs for visually grounded instruction tuning derived and augmented from 27 existing datasets. Unlike previous datasets that depend on bounding boxes or align only noun phrases (e.g. entity object), our framework allows the alignment of various crucial word types, including adjectives and verbs (Table 1).

Evaluating Visual Grounding. A variety of benchmarks have been developed to ensure that VLMs rely on visual content rather than textual

biases. Traditional methods such as VQA-CP (Agrawal et al., 2018) and GQA (Hudson and Manning, 2019) modify data splits or balance question-answer distributions to penalize overreliance on language priors, while synthetic sets like CLEVR (Johnson et al., 2017) remove commonsense priors to force explicit visual reasoning. Recent approaches (POPE (Li et al., 2023b), NaturalBench (Li et al., 2024)) introduce adversarial or carefully constructed examples that only a visually grounded model can solve. Others evaluate model robustness using image or question perturbations, such as CSS (Chen et al., 2020) that generates counterfactual samples by removing relevant nouns in the image or question and assigning new ground-truth answers, and CARETS (Jimenez et al., 2022) which blurs or masks irrelevant background regions to evaluate model consistency. In contrast to these perturbation-based methods, which often rely on annotated datasets and complex object selection, FiVL is applicable to any dataset and complements related benchmarks such as FiVL-POPE. FiVL explicitly identifies key visual expressions in a question-answer pair, applies vision masks, to compute a Visual Reliance Score. This metric assesses both the model reliance on the image as well as how well a benchmark necessitates visual context for accurate question answering.

3 FiVL Framework

In this section, we introduce our proposed framework, which offers two advantages over existing grounding datasets. Specifically, 1) our framework can augment any image-text dataset without relying on bounding box annotations, as these are generated on the fly, and 2) it enables fine-grained alignment with diverse types of textual content, extending beyond object entities as in prior work. We will then describe how our framework is utilized to generate both training and evaluation datasets.

3.1 Data Collection Pipeline

We built grounded datasets for training and evaluation, by enhancing vision-question-answer and instruction datasets. Figure 1 shows an overview of the pipeline. Each sample in the original datasets was augmented with key expressions, along with their corresponding bounding box indices and segmentation masks within the images as follows:

Key Expression Identification. The initial stage of data collection focused on identifying key expressions within each question-answer pair, using GPT-4o. We refer to key expressions as specific words or phrases, like object names, attributes, or spatial relations, that rely on the visual context provided by the image. We prompted GPT-4o with only the text of the question-answer pairs, omitting the images and asked it to detect essential expressions. The prompt is shared in Appendix 13. Using only questions and answers without visual cues allows GPT-4o to rely solely on linguistic context to determine whether certain words could be evoked based on text alone. This approach can help filter language-based answers from those needing visual context, while being computationally efficient. This process yielded a robust set of expressions, capturing the elements in each conversation that are closely tied to the visual information.

Bounding Box and Segmentation Masks. To accurately associate key expressions with specific regions in each image, we used the GroundedSAM pipeline (Ren et al., 2024), which employs the GroundingDINO-tiny model (Liu et al., 2024d) for initial expressions localization generating bounding box indices, followed by the Segment Anything vit-huge model (Kirillov et al., 2023) for precise segmentation mask creation. Each key expressions was mapped to its relevant visual region, creating high-quality segmentation maps. If multiple segments corresponded to a single phrase, they were consolidated into a unified mask assigned to each token within the phrase, to maintain consistency across annotations. We removed segmentation mask of the same sample that overlapped by more than 95%, ensuring that each segmentation map uniquely represents essential visual regions, avoiding redundancy and improving annotation clarity.

3.2 Training Dataset

Our training dataset, FiVL-Instruct, is built upon the LLaVA-1.5-mix-665K instruction tuning dataset (Liu et al., 2024b), a vision-language instruction dataset containing 665K structured conversations between users and GPT. Most interactions begin with a user-provided image, followed by related questions, and GPT responses, each question-answer pair is referred as a turn.

We augmented the original LLaVA-1.5-mix-665K dataset by integrating the key expressions and their segmentation masks according to the pipeline outlined in Section 3.1. Not every FiVL-Instruct sample includes a key expression. For such cases,

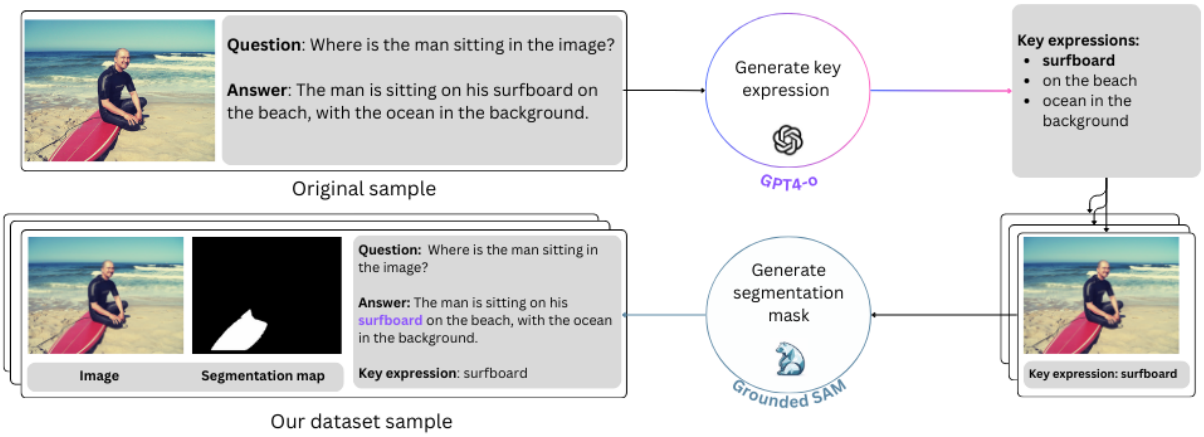


Figure 1: Dataset Collection Overview. First, GPT4-o processes the question and answer to produce "key expressions", which are then passed to GroundedSAM along with the image to produce segmentation maps.

we retained the original data point unchanged to maintain the dataset size for training. In our dataset, each conversation consists of multiple turns with an average of ten turns. Across the dataset, we collected 1.5 million unique segmentation masks for 2.3 million key expressions, averaging 2.3 masks and 3.5 key expressions per conversation. On average, a key expression consists of 2.4 words and the segmentation covers 28% of the image. We also analyzed the types of key expressions in the resulting dataset. As shown in Table 1, our expressions exhibit diverse types, making them distinct from those in prior grounding datasets.

Nouns	Adjectives	Proper Nouns	Adpositions	Verbs	Others
42%	14%	10%	9%	8%	17%

Table 1: Statistics FiVL-Instruct dataset, showing key expressions words types.

3.3 Evaluation Datasets

To assess the visual reliance of various LVLMs, we created three benchmark datasets derived from the following benchmarks: POPE (Li et al., 2023b), VQAv2 (Goyal et al., 2017), and GQA (Hudson and Manning, 2019).

We selected these benchmarks, because they each requires different levels of image reliance. POPE assesses sensitivity to visual perturbations, GQA evaluates understanding of detailed scene relationships, and VQAv2 tests visual grounding for diverse question types. Together, they offer a well-rounded assessment of how much models depend on visual information to answer accurately. We followed the procedure outlined in Section 3.1 and produced FiVL-POPE, FiVL-VQAv2, and FiVL-GQA datasets. To suit the nature of the evaluation datasets, we adapted the prompts for the key expression extraction (See Appendix B, Figure 14). Unlike FiVL-Instruct, we filtered out samples without key expressions or segmentation maps resulting in a reduction in dataset sizes. As a result, FiVL-POPE covers 65% of POPE, FiVL-VQA-v2 retains 40% of VQA-v2 and FiVL-GQA accounts for 95% of GQA size (refer Table 4 in Appendix for the actual size of the filtered dataset numbers). This indicates that the original GQA relies more on visual context than POPE and VQAv2. Our evaluation sets: FiVL-POPE, FiVL-VQAv2, and FiVL-GQA, select subsets from the original datasets that require visual context and are better suited for visual-alignment testing. Table 5 from Appendix presents additional statistics for these datasets.

4 Method Evaluation

To ensure the quality of our framework, we conducted a multi-step evaluation process on the training dataset described in Section 3.2. This included both human-based evaluations and automated assessments, allowing us to validate the relevance and accuracy of the key tokens and their alignment with visual content. Below, we outline the key components of our evaluation strategy.

4.1 Human Evaluation

We conducted a manual evaluation in order to validate the coherency of the key expressions as well as the relevancy of the segmentation maps with respect to the formers. For each sample, we presented to the annotators one random key expression with its associated segmentation map. Annotators were

asked three questions: *whether the key expression aligns with the definition provided in Section 3.1*, *if the segmentation map is relevant to the key expression*, and *whether the sample is of good quality* (does the text makes sense, is the answer related to the question). In total, 557 unique samples were annotated by 12 different annotators. A screenshot of the API is shown in Appendix H. Results show that 77% of the annotators labeled the samples as overall good data points. In the key expression evaluation, 75% of key expressions were deemed pertinent. In the segmentation map evaluation, 58% of segmentation map were annotated as relevant to the key expression. The last result can be explained by the fact that some key expressions might inherently be more abstract or complex or by the performance of the GroundedSAM pipeline. Finally, if we compute the key expressions and segmentations score only for the samples annotated as "good data point overall", 85% of the data are with valid key expressions and 69 % are with relevant segmentation masks. Additionnaly, we find that the quality of the segmentation is related to its size. Figure 6 from Appendix indicates that when the segmented mask occupies less than 20% of the image, annotators were more likely to consider the segmentation relevant. To train our model (see Section 5.1), we selected the segmentation masks based on their size. This metric can be used as a threshold-based filtering method for future applications of the dataset.

4.2 Automatic Evaluation

Inspired by recent applications using GPT4-o as-a-judge (Zheng et al., 2023a), we designed two prompting techniques to automatically assess the quality of extracted keywords and segmentation masks based on a given keyword. Both evaluations were conducted on a randomly sampled set of 1,957 keywords and their corresponding segmentations from FiVL-Instruct.

4.2.1 Keyword Evaluation

We prompt GPT4-o, prompts are presented in Appendix C, to evaluate the correctness of the key expressions and report the following metrics: *Importance Ratio = 76%* representing the percentage of extracted expressions classified as key expressions. This result is close to human evaluation, which is 75%. *Overall Importance Degree = 6.8*, which indicates the average importance score across all keywords, regardless of GPT-4o classification. And, *Importance Degree of Important Keywords = 9.0*, which calculates the average importance ratio of keywords identified as important by GPT-4o. These metrics indicate the high quality of our keywords.

4.2.2 Segmentation Evaluation

Given a keyword, we aim to evaluate whether our segmentation for this keyword is accurate. We designed two prompts to assess the quality of the segmentation: first, we check if the segmentation content adequately covers the keyword (Seg1); second, we verify that the inverse of the segmentation does not contain any content related to the keyword (Seg2). Both prompts are given in Appendix C. Results show that only for $Seg1 = 46\%$ of the cases GPT-4o capture the keywords in the segmentation. On one hand, this result aligns with the manual annotations and can be addressed in the same manner. On the other hand, we found that segmentations classified as good often involve specific objects (e.g., tennis players, bears). In contrast, segmentations classified as bad are often abstract concepts (e.g., water pressure, mental game, splashing), descriptive words (e.g., unique, uneven ground), or complex actions (e.g., walking over logs). These types of words are difficult to link to a specific part of an image when the full image context is not provided. This also highlights the limitations of the first type of evaluation prompt. In $Seg2 = 72\%$ of cases, the model determines that the inverse of the segmentation is irrelevant to the keywords, accurately recognizing that without the segmented mask, the key expressions are not present in the image. This measures if we do not miss key objects in our segmentation maps. If 2 objects appear in the image not at the same positions, we make sure that our maps contain both of them.

5 Applications of FiVL Datasets

In this section, we describe three approaches to utilize our datasets. Section 5.1 describes how FiVL can be used as a training dataset and the resulting models not only achieve better performance but also has one more capability than the baseline model: generate segmentation maps. Section 5.2 introduces FiVL as a tool for evaluating the visual reliance of LVLMs. Section 5.3 shows that FiVL can assist the interpretability of models.

5.1 Training

We introduce here a training task referred to as Vision Modeling. To assess the effectiveness of this

task, we fine-tuned an LVLM, specifically, LLaVA-1.5-7b (Liu et al., 2023), referred as to the baseline, on FiVL-Instruct. For training our model, we used only key expressions that appeared verbatim in the answers for each turn, focusing exclusively on noun-based key expressions.

Method. In the original LLaVA training, it has two stages: the first pretraining stage trains a projector which aims to align visual and textual representations, while the second finetuning stage performs only language modeling on the textual outputs of the LM head. In this work, we propose to guide the visual outputs of the last linear layer during the finetuning stage, in addition to performing language modeling on its textual outputs. Our approach augments the Language modeling cross-entropy loss with a Vision modeling (VM) cross-entropy loss where each patch that belongs to a segmentation map is trained to predict the related keyword from the vocabulary.

We denote by x the input and y the logits with respect to each token. The logits are the outputs of the last linear layer that projects the last hidden states to the vocabulary space:

$$x = (x_{i_0}, x_{i_1}, \ldots, x_{i_{N_i}}, x_{t_0}, x_{t_1}, \ldots, x_{t_N}),$$
$$y = (y_{i_0}, y_{i_1}, \ldots, y_{i_{N_i}}, y_{t_0}, y_{t_1}, \ldots, y_{t_N}),$$

where N_i is the number of image tokens, N_t the number of text tokens, N the total lenght. x_i are the inputs embedding that relate to the image tokens and x_t to the text tokens; $y_i \in \mathbb{R}^{N_i \times \text{vocabulary\_size}}$ represents visual logits, while $y_t \in \mathbb{R}^{N_t \times \text{vocabulary\_size}}$ represents textual logits.

In Language Modeling (LM), only y_t related to the answer are trained. We propose to also train y_i related to the segmented piece. Figure 2 shows an example where given a picture, a question, the LM loss would only guide the relevant tokens y_t to be the expected answer *The man is sitting on his surfboard* <...>. In our method, we also do vision modeling by training each visual logit corresponding to the segmented mask to refer to the noun from the key expression: *surfboard* from the text vocabulary. In order to create the vision labels we proceed like such: for each sample, each image token will be assigned to exactly one token in the text vocabulary. The selection is based on the size of the mask (we take the smallest) and the type of the keyword (we filter only nouns). That way, for each image patch, there is maximum one key token that

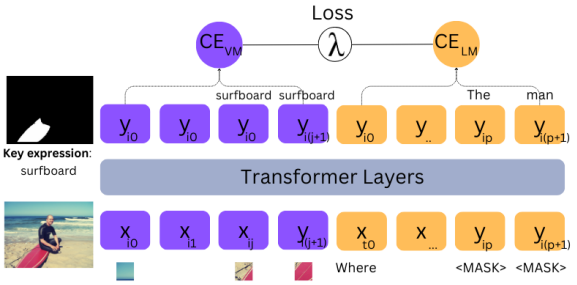


Figure 2: Overview of Vision Modeling pretraining task.

describes the patch. Image patches that do not have a related keytoken are ignored in the loss, similar to LM. We then compute a weighted sum from the cross-entropy, CE_{VM} between the created vision labels and the visual logits and the cross-entropy related to language modeling, CE_{LM}. The resulting loss is computed as such:

$$L = \lambda * CE_{VM} + (1-\lambda) * CE_{LM}, \lambda \in [0,1] \quad (1)$$

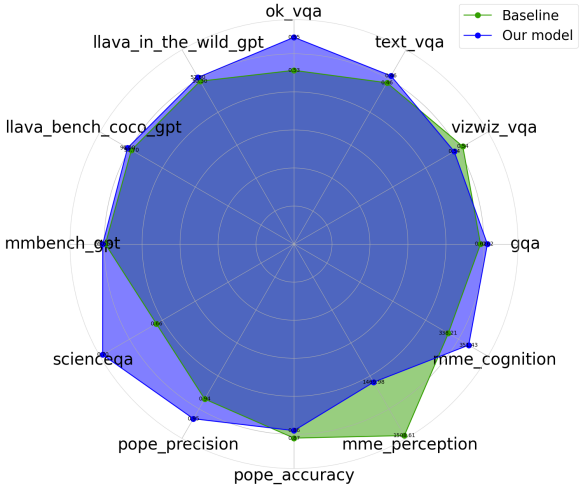


Figure 3: Our model trained on FiVL-Instruct evaluated on various benchmarks compared to the baseline.

Improve Benchmark Results. We conducted multiple experiments to determine the optimal hyperparameters. We finetuned LLaVa-v1.5-7b from scratch using our augmented dataset. We used the trained multimodal projector and started from Vicuna-v1.5-7b (Zheng et al., 2023b) weights. We maintained the original training setup (batch size, number of epochs, etc.) and primarily focused on experimenting with different learning rates and λ. The best results were achieved with a learning rate of 2-e5, the same as in the original setup, and λ set

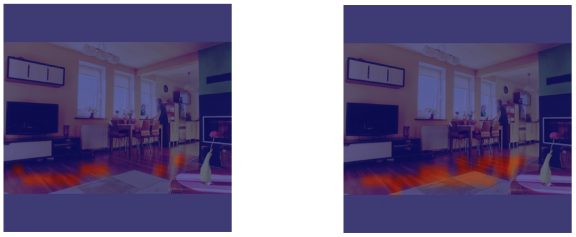
to 0.1. Training details are shared in Appendix E and ablations are reported in Appendix F. Figure 3 shows how we outperformed the baseline in different benchmarks: OK-VQA (Marino et al., 2019), MME (Fu et al., 2024), POPE (Li et al., 2023b), ScienceQA (Lu et al., 2022), MMBench (Liu et al., 2024e), LLaVA-Bench-COCO (Liu et al., 2023), LLaVA-in-the-wild (Liu et al., 2023), Text-VQA (Singh et al., 2019), VizWiz-VQA (Gurari et al., 2018), GQA (Hudson and Manning, 2019).

Better Grounding Outcome. Figure 4 compares the baseline model and FiVL, illustrating the correspondance of each image patch with its related most probable token from the vocabulary. The argmax of the vision logits is identified, mapped back to the text vocabulary. Then, for each token, the relevant image patches are highlighted, indicating which parts of the image align with that token. Although the baseline can capture some relevant text tokens for the image patches and tends to scatter semantically similar image patches across different tokens from the vocabulary. Some of these tokens may be relevant, but others are not, indicating a lack of consistent grounding. On the other hand, our model shows more relevant images patches related to the word.

Vision Logits as Approximate Segmentation Maps. Another interesting finding is that we can obtain a weak “segmentation maps" by predicting the most probable text tokens from the vision logits. As a simple observation, averaging over 100 examples, the baseline predicts 74 different tokens overall (with lots of unrelated tokens such as "a", "*", "is" etc.), while our model only encompasses 9 tokens. This demonstrates potential in leveraging visual logits for segmentation. as shown in Figure 4. More examples are presented in Figure 12 in Appendix G. We further conducted evaluations to assess the performance of the segmentation capability of our model. Results, reported in Appendix G, shows FiVL's enhanced ability to produce precise and coherent segmentation masks.

5.2 Visual Reliance Evaluation

FiVL datasets also allow us to measure *Visual Reliance* by performing perturbation based evaluation: first assessing model accuracy on the original images, then on the masked images. We introduce a *Visual Reliance Score* in Eq.2, which measures the percentage of drop in accuracy from the original to the masked image version. A higher score indicates stronger model dependency on visual input.



(a) Baseline (b) Our model

Figure 4: Predicted token from vision logits ("Flo", for "floor") and its corresponding regions in the image.

$$\text{Visual Reliance Score} = \frac{\text{accuracy}_{\text{original}} - \text{accuracy}_{\text{perturb}}}{\text{accuracy}_{\text{original}}} \quad (2)$$

Indeed, the perturbation based on the masked image is not perfect, it still provides a measurement of visual reliance. To confirm that FiVL is suitable for evaluating visual reliance, we created a control dataset with random masking. In this control set, each image contains a bounding box mask of the same size as the key expression mask, but placed at a random location within the image. This approach provides a comparison to determine whether performance declines specifically due to masking critical visual areas or simply from general occlusion.

We compared the performance of two models, LLaVA-v1.5-13b and Qwen2-VL-7B-Instruct (Wang et al., 2024), across the three evaluation datasets we created.

Compare Perturbation Methods. Table 2 compares FiVL and Random Perturbation. It shows that across all benchmarks and models, the perturbation based on FiVL masks causes a significantly larger performance drop compared to random perturbation. This indicates that our bounding boxes capture meaningful visual content relevant to the questions and FiVL represent good testbeds for visual reliance.

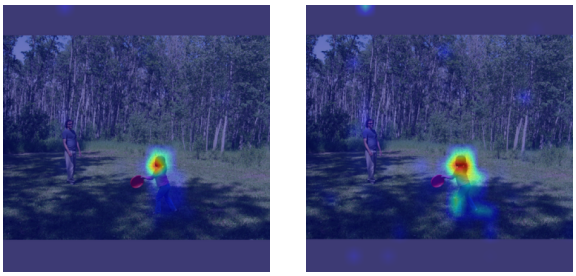
	VQA-v2		GQA		POPE	
	FiVL	Random	FiVL	Random	FiVL	Random
LLaVA-13B	0.72	-0.05	0.33	0.03	0.49	0.02
Qwen2-VL-7B	0.64	0.07	0.38	0.03	0.47	0.02

Table 2: Comparison of Visual Reliance Score between FiVL bounding boxes and random perturbations across benchmarks and models.

Compare Models and Benchmarks. To gain a broader understanding of model/benchmark performance, we evaluated five models on FiVL-VQAv2, FiVL-POPE, and FiVL-GQA. This helps to assess the generalizability of our approach across more models. Table 3 shows our results for Qwen2-VL-7B, LLaVA-v1.5-7b(Liu et al., 2023), LLaVA-13B, GPT4o (OpenAI, 2024), BLIP-2 (Li et al., 2023a), Pixtral-12B (Agrawal et al., 2024) and Phi3-Vision(Abdin et al., 2024), which are state-of-the arts mutlimodal models. In bold, are the highest visual reliance scores per model and across all benchmarks. The results unanimously indicate that, among all models, FiVL-VQAv2 requires models to rely on the image the most compared to other datasets. Underlined are the highest visual reliance scores across models, given a benchmark. Looking at the average performance per model across benchmarks (last column), we observe that GPT4-o relies most heavily on the image as a reference for answering, followed by Pixtral-12B. Lastly, we observe a correlation between overall model performance and the Visual Reliance score. According to the available VLM Leaderboard (OpenCompass, 2025), that measures the performance of the models on a broad range of benchmarks, and Table 3, we see that GPT4o (ranked 20 on the leaderboard) has a higher overall Visual Reliance Score compared to Pixtral-12B (ranked 54). Within a similar Visual Reliance Score range, follow LLaVA-13B (118), Qwen2-VL-7b (136) and Phi3-V (77). Lastly, LLaVA-7b (127) appears at a lower rank. This suggests that this average of Visual Reliance Scores captures the overall performance of the model and is not overly sensitive to the specific benchmarks used. All together, these results indicate that effective image utilization is a key factor in achieving higher performance.

	VQA-v2	GQA	POPE	Avg VRS
Qwen2-VL-7B	**0.64**	0.38	0.47	0.50
LLaVA-13B	**0.72**	0.33	0.49	0.51
LLaVA-7B	**0.56**	0.31	0.47	0.45
GPT4o	**0.74**	0.63	0.49	0.62
BLIP-2	**0.52**	0.23	0.03	0.26
Pixtral-12B	**0.75**	0.58	0.42	0.58
Phi3-V	**0.60**	0.33	0.54	0.49
Avg	**0.65**	0.40	0.42	-

Table 3: Visual reliance scores (VRS): % of drop in performance using FiVL bounding boxes for perturbations. In **bold**: highest scores across benchmarks. Underlined: highest scores across models. Avg stands for Average.



(a) Attention Head (10,6) (b) Attention Head (14,11)

Figure 5: Attention heatmaps overlaid on the original images for attention heads (10,6) and (14,11) of the token "girl" for the answer: *The two people [...] are a man and a little **girl**.*

5.3 Explainability

We show that FiVL can assist the interpretability of LVLMs by generating a summary plot showing a vision-alignment metric computed across all heads and layers, as introduced in (Aflalo et al., 2022). Using Spearman correlation between the segmentation mask of FiVL-Instruct dataset and the attention to the corresponding key expression tokens in the Vision-to-Language attention component, we are able to retrieve the heads achieving the strongest VL alignment. The head summary (Appendix D, Figure 9) indicates that heads (10,6) and (14,11) are effective at aligning vision with language. For instance, Figures 5a and 5b show from which patches of the image the token *girl* gets the most attention, clearly focusing on the girl.

6 Conclusion

In this paper, we introduced FiVL, a framework designed to enhance vision-language alignment in large vision-language models. We applied our approach across key stages of an LVLM training workflow: training, evaluation, and explainability. By training a LLaVA model using the FiVL dataset and our novel training task, we measured improvement in a majority of benchmarks and produced a built-in feature that segments the image. Our evaluation datasets measured model reliance on images for answering questions, offering insights into the level of image dependency required across benchmarks. The results indicate a correlation between this dependency and overall model performance. Finally, our explainability application enables users to identify attention heads that excel in vision-language alignment, allowing for a deeper understanding of potential hallucinations.

Limitation

In this work, we utilize the FiVL framework to augment LLaVA instruction fine-tuning data and train a new model to compare against the baseline, demonstrating the effectiveness of our proposed framework and training objectives. However, we have only investigated LLaVA model, because of the limited availability of open-source training datasets for other LVLMs and augmenting additional data incurs additional inference costs. Lastly, we rely on an off-the-shelf segmentation model (GroundedSAM) that takes a simple text prompt and an image as input. In our context, this may lead to less accurate segmentation, as the full contextual understanding of keywords might be necessary. To mitigate this issue, we could apply a filtering technique to enhance the overall quality of the dataset.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. *Preprint*, arXiv:2203.17247.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. *Preprint*, arXiv:1712.00377.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. *Preprint*, arXiv:2003.06576.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *Preprint*, arXiv:1802.08218.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Carlos E. Jimenez, Olga Russakovsky, and Karthik Narasimhan. 2022. Carets: A consistency and robustness evaluative test suite for vqa. *Preprint*, arXiv:2203.07613.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

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

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Preprint*, arXiv:2410.14669.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024d. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Preprint*, arXiv:2303.05499.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024e. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. *Preprint*, arXiv:1906.00067.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenCompass. 2025. Open vlm leaderboard. `https://huggingface.co/spaces/opencompass/open_vlm_leaderboard`. Accessed: 12/02/2025.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. 2024. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13009–13018.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *Preprint*, arXiv:2401.14159.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

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. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. 2024a. Llava-Grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 2024b. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14227–14238.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

A Appendix

Appendix

A Evaluation dataset

Table 4 compares the size of our datasets with the original datasets and Table 5 presents some statistics of FiVL-VQA-v2, FiVL-GQA and FiVL-POPE.

	VQA-v2	GQA	POPE
Original	9,999	12,280	9,000
FiVL	4,040	11,660	5,870

Table 4: Evaluation dataset sizes after filtering out samples without key expressions or segmentation masks.

	FiVL-VQAv2	FiVL-GQA	FiVL-POPE
Key expressions	1.27	1.5	1
Segmentation masks	3.79	4.71	3.48
% of masked pixels	24%	21%	16%

Table 5: Statistics per sample of our evaluation datasets. First row details the average number of key expressions, second row describes the average number of distinct segmentation masks and last row describes the average percentage of the pixels that were masked.

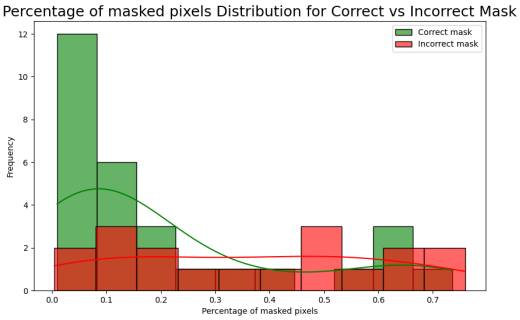


Figure 6: Impact of the size of the segmentation mask. Comparison of the Percentage of masked pixels Distributions for Correctly and Incorrectly annotated Masks

B System prompts for key expressions retrieval

We use GPT-4o via the Azure OpenAI API to extract the key expressions of the datasets we considered. In this section, we share the prompts used for this step of the data collection. We had to use slightly different prompts for the training datasets compared to the evaluation datasets. In the training datasets, where instructions are open-ended question-answer pairs, the key expressions are often found in the answer. However, in the evaluation datasets, we encountered questions that required specific types of responses (yes/no questions, counting etc...). In these cases, the key expressions are typically found in the question instead. For references, we have provided the prompt used for training dataset in Figure 13 and prompts used for evaluation datasets VQA-V2, GQA, and POPE in Figure 14. For each benchmark we use different examples that suit the best to the types of questions. See Figure 15 for FiVL-VQAv2 and Figure 16 for FiVL-GQA and FiVL-POPE.

C System prompts for evaluation

We used GPT-4o as LLM-as-a-judge in order to evaluation the correctness of the key expressions and the segmentation maps. The system prompt are shared in Figure 7 and Figure 8.

```
[Seg1] You are given a part of the image
 and a word/phrase, do you think this is
 a good segmentation that the given part
 of the image covers this word/phrase?

Word/phrase: {word}

Answer only "yes" or "no".

[Seg2] You are given a part of the image
 and a word/phrase, do you see any part
of the image that is related to the word
?

Word/phrase: {word}

Answer only "yes" or "no".
```

Figure 7: Segmentation Verification Prompt for GPT-4o.

```
You are given a question, a word/phrase
and an image. Please rate the importance
 degree from 0-10 scale ([OID]).
Note that
 - 0 means not important at all and 10
means very important.
 - Important word/phrase means that this
 word/phrase is closely related to the
image and the question, and it could not
 be evoked without the use of the image
(IR).
 - If the question does not related to
the image, in other words, the answer
does not depend on the image content,
then any words are not important.

Question: {question}

A word: {word}

Only answer important or not important,
and the importance degree from 0-10?
```

Figure 8: Keyword Verification Prompt for GPT-4o.

D Explainability

For an attention matrix of size $(N_{layers}, N_{heads}, N_i + N_t, N_i + N_t)$, The head summary calculates the statistical mean over the last two dimensions, producing a plot with dimensions of (N_{layers}, N_{heads}) averaged for 500 samples. For a given question, image, key expression and related segmentation mask from the FiVL-Instruct dataset, we generate the answer using LlaVA-v1.5-7b. We then identify if the key expression is in the answer or in the question. If so, we probe each head by computing the Spearman correlation between the segmentation mask $(\sqrt{N_i}, \sqrt{N_i})$. and the attention to the corresponding key expression tokens in the Vision-to-Language attention component $(1, 1, N_i, 1)$ (first dimension selects the layer, second the head and the last dimension corresponds to the key token) for each head. This is performed on the language model component but not on the vision component of LLaVA. In this way, we identify the attention heads that ground the most the two modalities by performing a function similar to object segmentation. Figure 9 shows the head summary and the corresponding language-vision attention weights related to the key expression tokens displayed as a heatmap over the image. The head summary shows that the heads achieving the strongest vision-language alignment are in the early layers. This might be due to the fact that the input to this transformer is the output of multimodal projector of LLaVA, which is designed specifically to align these two modalities. The head summary indicates that heads $(10, 6)$ and $(14, 11)$ are effective at aligning vision with language. In this way, we identify the attention heads that ground the most the two modalities by performing a function similar to object segmentation. Figure 9 shows the head summary and the corresponding language-vision attention weights related to the key expression tokens displayed as a heatmap over the image. The head summary shows that the heads achieving the strongest vision-language alignment are in the early layers. This might be due to the fact that the input to this transformer is the output of multimodal projector of LLaVA, which is designed specifically to align these two modalities. The head summary indicates that heads (10,6) and (14,11) are effective at aligning vision with language.

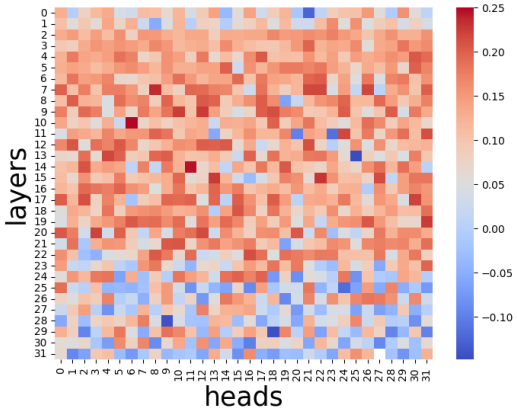
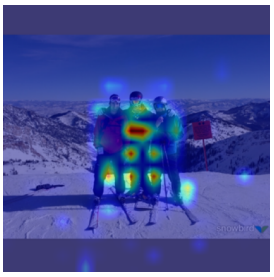


Figure 9: Head summary for VL alignment via Spearman correlation between token segmentation and vision attention

(a) Attention Head (10,6) of the token *three*. A - *There are **three** people in the image.*



(b) Attention Head (14,11) of the token *three*. A - *There are **three** people in the image.*

Figure 10: Attention heatmaps overlaid on the original images for attention heads (10,6) and (14,11), which have a high Spearman correlation, to probe vision-language alignment.

E Training details

To train our model, we used 8 Nvidia RTX A6000 GPUs using the hyperparameters from Table 6

Batch Size	4
Number of GPUs	8
Gradient Accumulation	4
Number of epochs	1
LLaVA Image Size	576
Optimizer	SGD
Learning Rate	$2e-5$
λ_{VM}	0.1
BF16	$True$
LR scheduler	$cosine$
Vision Tower	openai/clip-vit-large-pa
Language Model	lmsys/vicuna-7b-v1.5

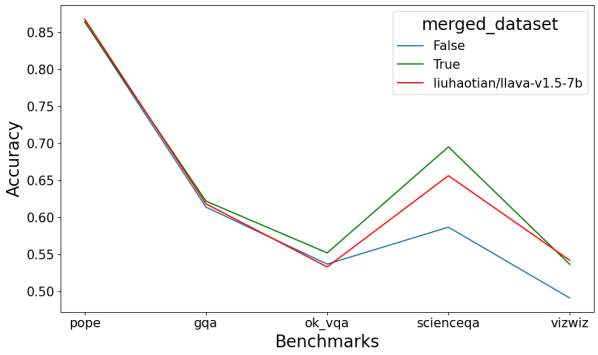
Table 6: Hyperparameters to train our model.

F Ablations

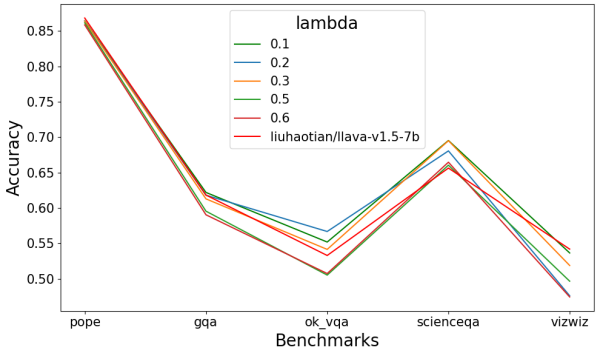
We conducted ablations studies on different parameters of the model. In this section, we will limit the experiments on a subset of the benchmarks. As mentioned in Section 3, the FiVL-Instruct dataset includes some samples without key expressions. We first trained our model using only on the samples that had at least one associated key expression and segmentation map. We conducted another experiment by merging the remaining samples with these. Figure 11a presents the results of this ablation, showing that merging the samples lead to improved performance, even over the baseline LLaVA-v1.5-7B model.

The second ablation focused on the λ parameter, which controls the weight of the vision modeling loss, as outlined in Section 5.1 and equation 1. The optimal performance was obtained with $\lambda = 0.1$. As shown in Figure 11b, our approach also outperforms the baseline for all $\lambda \leq 0.3$.

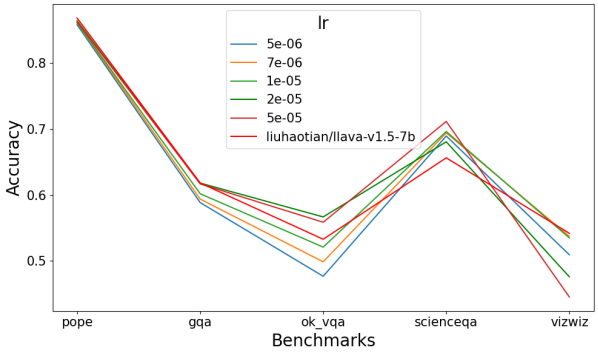
Finally, since we are introducing a new capability in the training, we experimented with different learning rates to see if it would lead to improved convergence or better overall performance. Figure 11c shows that overall the original learning rate of $2e-5$ achieved the best performance.



(a) Ablations on the dataset size



(b) Ablations on the λ parameter



(c) Ablations on the learning rate

Figure 11: Ablations for the training method

G Performance of the segmentation maps inherently provided by our model

To evaluate the segmentation ability of our FiVL model, we evaluated Intersection-Over-Union

(IoU) on a subset of 10,000 images from the GQA-val dataset. For each sample, we perform an inference using the baseline LLaVA-7b and our model. From the outputs, we retrieve the visual logits for each visual token, we assigned a text token from the vocabulary corresponding to the maximum logit probability, referred to as the max-v token. By aggregating all image tokens associated with each max-v token, we effectively generated a segmentation mask for each represented text token, like describe in Section 5.1. Additionally, as ground truth to compare against, we employed Grounded-SAM to produce segmentation maps given each max-v token. Grounded-SAM was implemented using the IDEA-Research/grounding-Dino-Tiny model with thresholds set at 0.2, 0.4, and 0.6, followed by facebook/sam-vit-huge with a threshold of 0.0. The Intersection over Union (IoU) score was computed between the FiVL-generated segmentation masks and the corresponding Grounded-SAM masks to quantitatively assess alignment. To provide a comparative analysis, we also computed IoU scores for the segmentation masks produced by the baseline model. As detailed in Table 7, across all thresholds, FiVL generated approximately 7 times fewer max-v tokens per image compared to the baseline model (column #tokens/sample), indicating more concise and semantically meaningful segmentation. FiVL also showed significant improvement in average IoU scores (column IoU), increasing approximately three times: from 0.05 to 0.18 at a threshold of 0.2, from 0.06 to 0.21 at 0.4, and from 0.09 to 0.24 at 0.6, showcasing its superior ability to generate precise and coherent segmentation masks. In general, across all thresholds, the baseline generates significantly more max-v tokens per image, resulting in a higher number of samples with segmentation maps found by Grounded-SAM (column #samples). Finally, the percentage of tokens processed by Grounded-SAM is substantially higher for our model compared to the baseline (column #processed), indicating that the max-v tokens retrieved by our model were more meaningful than those from the baseline. Figure 12 shows the segmentation maps we obtained for the max-v token describing each image. For example for the example 12a, we computed the argmax of the tokens highlighted in red, and it corresponded to the token "bear" in the vocabulary

	Thresh	IoU	# tokens/sample	# samples	#processed
Baseline	0.2	0.05	73.3	10,000	0.89
Our Model		**0.18**	10.3	10,000	0.96
Baseline	0.4	0.06	73.3	10,000	0.40
Our Model		**0.21**	10.3	9,983	0.65
Baseline	0.6	0.09	73.4	9,326	0.08
Our Model		**0.24**	10.6	8,604	0.26

Table 7: Performance of the segmentation maps inherently provided by our model

H API of the manual evaluation

Figure 17 shows the API used for the manual evaluation done on FiVL-Instruct. Given a question, an answer (on the left) and an image with a segmentation mask (on the right), the annotator had to answer the 3 following yes/no questions: is { key expression } correctly represented in the mask? Is {key expression} a significant word in the answer? Is this example generally good to be included in the dataset?

(a) Bear

(b) Bird

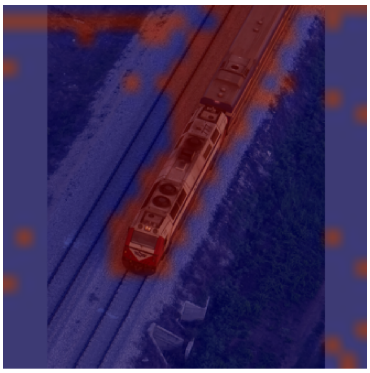
(c) Birds

(d) Bott

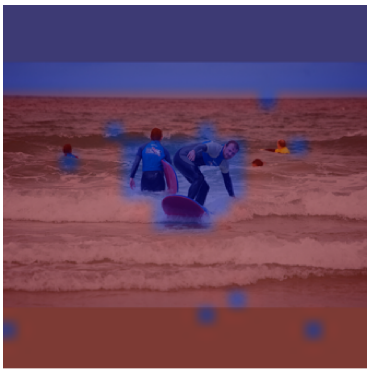
(e) Chair

(f) Dog

(g) People



(h) Train



(i) Water

Figure 12: Segmentations produced inherently by our model. Each figure corresponds to the max-v token specified in caption. Max-v token being the token realizing the maximum for each highlighted patch

FiVL-Instruct system prompt

A multimodal instruction-following dataset used for visual instruction tuning and it contains an image and a conversation. The conversation is constructed from a few turns of questions and answers regarding the image.
Given only a question and answer pair: identify short expressions from the answer which could not be generated without the image.
The expression

- expresses a visual content from the image.
- should be as short as possible.
- should not be longer than 4 words
- should not include punctuation
- should no include reference to the image

Unrelated expressions should be separated by the following string: ":::"
Don't add any additional information to the prompt.
For example:
Q: What are the giraffes doing in the image? <image>
A: The baby giraffe is walking next to the mother giraffe, both moving through the open area of their enclosure
The output should be as following:
baby giraffe:::mother giraffe :::open area of their enclosure
Identify the tokens for the following:
Q: {question}
A: {answer}

Figure 13: FiVL-Instruct system prompt used for training datasets

FiVL system prompts for the evaluation datasets

A multimodal instruction-following dataset used for visual instruction tuning and it contains an image and a conversation. The conversation is constructed from a few turns of questions and answers regarding the image.
Given only a question and answer pair: identify short expressions from the answer or the question which could not be generated without the image.
The expression

- should hypothetically express an immediate visual content from image. Thus, yes/no is NOT an expected expression, and some pronouns like "this", "that", "there", and "those" are not expected expressions.
- should be as short as possible.
- should not be longer than 4 words.
- should not include punctuations.
- should not include reference to the image, like: "the man wearing a blue suit in the image" it should be: "man wearing blue suit".

Unrelated expressions should be separated by the following string: ":::".
If there is no expressions found, your response must be "N/A".
Do not add any additional information to the prompt.

<EXAMPLES>

Identify the expressions for the following:
Q: "{question}"
A: "{answer}"

Figure 14: FiVL-Instruct system prompt used for the evaluation datasets

FiVL-VQAv2 - Examples

Example 1:
Q: "Considering the limited space and packed fixtures, what suggestions can be provided for organizing the bathroom to optimize its use?"
A: "To optimize the use of the small bathroom with tightly packed fixtures, there are some organization solutions that can be recommended: Install wall-mounted storage: By using vertical space on the walls and areas above the toilet, you can create additional storage for toiletries and other bathroom essentials without taking up floor space. Utilize narrow shelves or over-the-door organizers: Placing narrow shelves or over-the-door organizers on available narrow spaces can help with the organization of toiletries and other small items. Use multi-functional items: Opt for items that serve more than one purpose, such as a toilet paper holder with a shelf or a mirror with built-in storage. Opt for compact accessories: Choose smaller or more compact versions of bathroom essentials, like toothbrush holders or soap dishes, to maximize space on countertops and around the sink. Declutter regularly: Frequently assess your bathroom supplies and remove any items that are not in use or have expired to keep the bathroom tidy and make the most of the limited space. By following these suggestions, the bathroom can appear less cluttered, and occupants can make better use of the available space."
Key Expressions: small bathroom:::tightly packed fixtures:::vertical space on the walls:::areas above the toilet:::over-the-door organizers:::narrow spaces:::space on countertops and around the sink

Example 2:
Q: "How many shadows are on the ground?"
A: "3"
Key Expressions:shadows:::3

Figure 15: Examples for FiVL-VQAv2

FiVL-GQA and FiVL-POPE - Examples

Example 1:
Q: "Considering the limited space and packed fixtures, what suggestions can be provided for organizing the bathroom to optimize its use?"
A: "To optimize the use of the small bathroom with tightly packed fixtures, there are some organization solutions that can be recommended: Install wall-mounted storage: By using vertical space on the walls and areas above the toilet, you can create additional storage for toiletries and other bathroom essentials without taking up floor space. Utilize narrow shelves or over-the-door organizers: Placing narrow shelves or over-the-door organizers on available narrow spaces can help with the organization of toiletries and other small items. Use multi-functional items: Opt for items that serve more than one purpose, such as a toilet paper holder with a shelf or a mirror with built-in storage. Opt for compact accessories: Choose smaller or more compact versions of bathroom essentials, like toothbrush holders or soap dishes, to maximize space on countertops and around the sink. Declutter regularly: Frequently assess your bathroom supplies and remove any items that are not in use or have expired to keep the bathroom tidy and make the most of the limited space. By following these suggestions, the bathroom can appear less cluttered, and occupants can make better use of the available space."
Key Expressions: small bathroom:::tightly packed fixtures:::vertical space on the walls:::areas above the toilet:::over-the-door organizers:::narrow spaces:::space on countertops and around the sink
Example 2:
Q: "Is there a snowboard in the image?"
A: "no"
Key Expressions: snowboard

Figure 16: Examples for GQA and POPE prompts

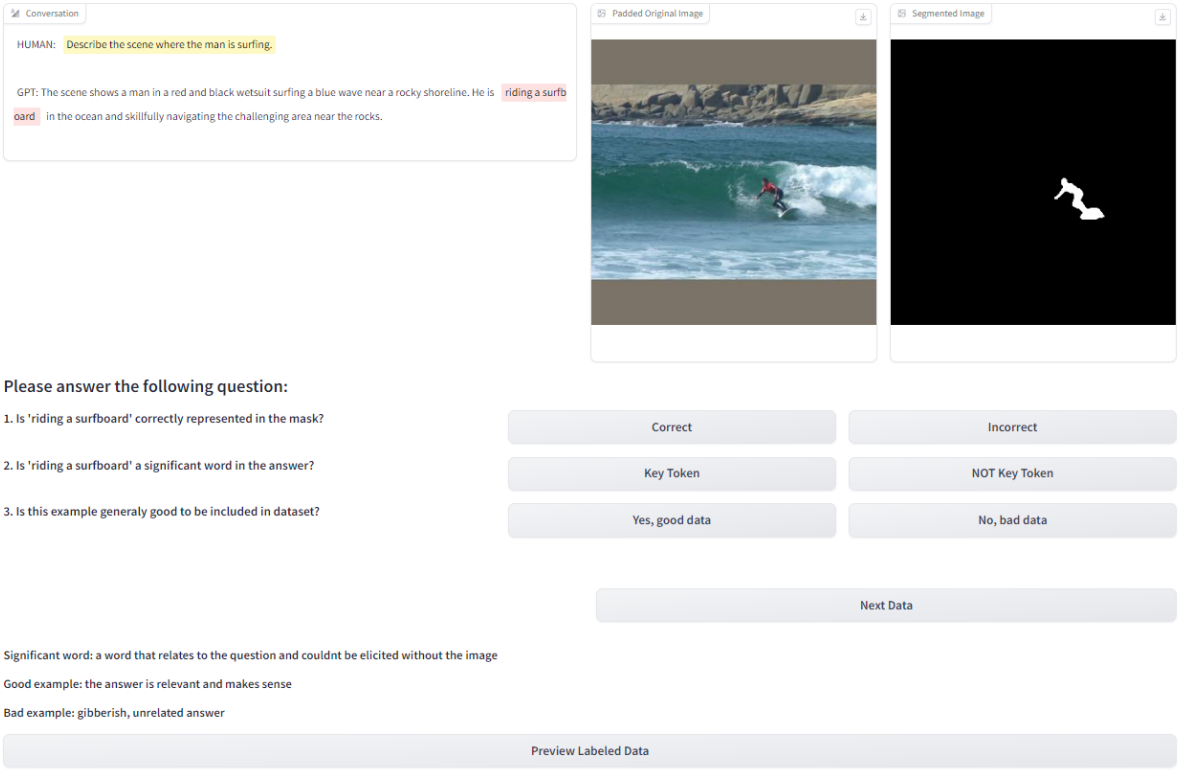


Figure 17: Web user interface for our dataset evaluation