FiVL: A Framework for Improved Vision-Language Alignment

Anonymous CVPR submission

Paper ID 12778

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. This issue extends to vision-language benchmarks, where it is difficult to make the image indispensable for accurate answer generation, particularly in vision question-answering tasks. In this work, we introduce FiVL, a novel method for constructing datasets designed to train LVLMs for enhanced visual grounding and to evaluate their effectiveness in achieving it. These datasets can be utilized for both training and assessing an LVLM's ability to use image content as substantive evidence rather than relying solely on linguistic priors, providing insights into the model's reliance on visual information. To demonstrate the utility of our dataset, we introduce an innovative training task that outperforms baselines alongside a validation method and application for explainability.

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 [18, 25, 26, 30]. 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 [19]. A common approach to mitigate this issue has been to introduce a visual grounding mechanism to the model [4, 26, 27, 30, 32, 33].

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 [4, 26, 30]. Bounding boxes, however, are coarse coordinates and unable to highlight objects or abstract concepts in finer detail. Additionally, the causal relationship between the generated bounding box and the response is not interpretable, and may instead be irrelevant to how the response is actually produced. Recent work have addressed these concerns by applying pixel-level grounding through the use of segmentation masks instead [27, 32, 33].

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 [23, 27, 32, 33]. To address these challenges and simplify the training of visually grounded LVLMs, we introduce a novel method, FiVL (Framework for Improved Vision-Language Alignment) for constructing datasets with visual-concept alignment. Different from prior methods, our framework leverages existing datasets by employing SoTA segmentation models while prioritizing visual information that is crucial for accuracy. The main contributions of this paper are as follows:

- We introduce a framework for augmenting existing datasets with segmentation masks that align visual regions and the corresponding text
- We propose a novel pretraining task that jointly trains text and vision tokens to generate grounding masks
- Utilizing the proposed dataset, we construct a framework based on perturbations to assess a model's dependency on visual input and the necessity of visual context for providing accurate responses
- Finally, we leverage the dataset for an application focused on explainability.

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 [25], LLaVA [18], Qwen2-VL [30], and many others [5, 6, 37], demonstrating advanced visual reasoning ability. Additionally, to link language output with visual input, 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 [26], Shikra [4], BuboGPT [34], Ferret [31], Qwen2-VL [30], and Groma [23]. To obtain a fine-grained localization of objects and semantic concepts pixel-level grounding has subsequently proposed in models such as Llava-Grounding [32], GLaMM [27], and GROUNDHOG [33].

Visually Grounded Datasets. Training LVLMs require large-scale visual instruction-following data [18]. 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 [23], a custom dataset, Groma Instruct, was constructed by prompting GPT-4V to generate grounded conversations based on 30K samples with region annotations from COCO [15] and VG [12]. Llava-Grounding [32] 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 [18] using GPT-4. The Grounding-anything Dataset (GranD) was specifically constructed to train GLaMM [27] 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 [33]. 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.

3. Our Framework

3.1. Data Collection Pipeline

We created grounded datasets for both training and evaluation, building upon existing vision-question-answer and instruction datasets. Each sample in the original datasets was augmented with key expressions, along with their corresponding bounding box indices and segmentation masks within the images. The data collection pipeline proceeded as follows:

Key Expression Retrieval. The initial stage of data collection focused on identifying key expressions within each question-answer pair, using GPT-4o. These expressions are specific words or phrases that would be unattainable without the visual context provided by the image, such as object names, attributes, or spatial relations. We provided only the text of the question-answer pairs, omitting the images, and prompted GPT-4o to detect essential expressions with a custom-designed prompt (Appendix A). 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 unique 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 [28], which employs the GroundingDINO-tiny model [20] for initial expressions localization generating bounding box indices, followed by the Segment Anything vit-huge model [11] 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. During this process, we filtered out redundant key phrases, retaining only unique tokens to enhance the dataset's precision. Additionally, if a segmentation mask overlapped by more than 95% with another mask in the same sample, only one of the masks was retained. This filtering step ensured that each segmentation map uniquely represented essential visual regions, avoiding unnecessary redundancy and improving annotation clarity. No synthetic images or simulations were used; instead, we enriched existing images with detailed annotations that strengthen the visual grounding of language.

3.2. Training Dataset

Our training dataset, FiVL-Instruct, is built upon the LLaVA-1.5-mix-665K instruction tuning dataset [16], a public vision-language instruction dataset containing 665K structured dialogues between users and GPT. Most interactions begin with a user-provided image, followed by questions related to the visual content, with GPT offering 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 accourding to the pipeline outlined in Section 3.1. Not every FiVL-Instruct sample includes a key expres-

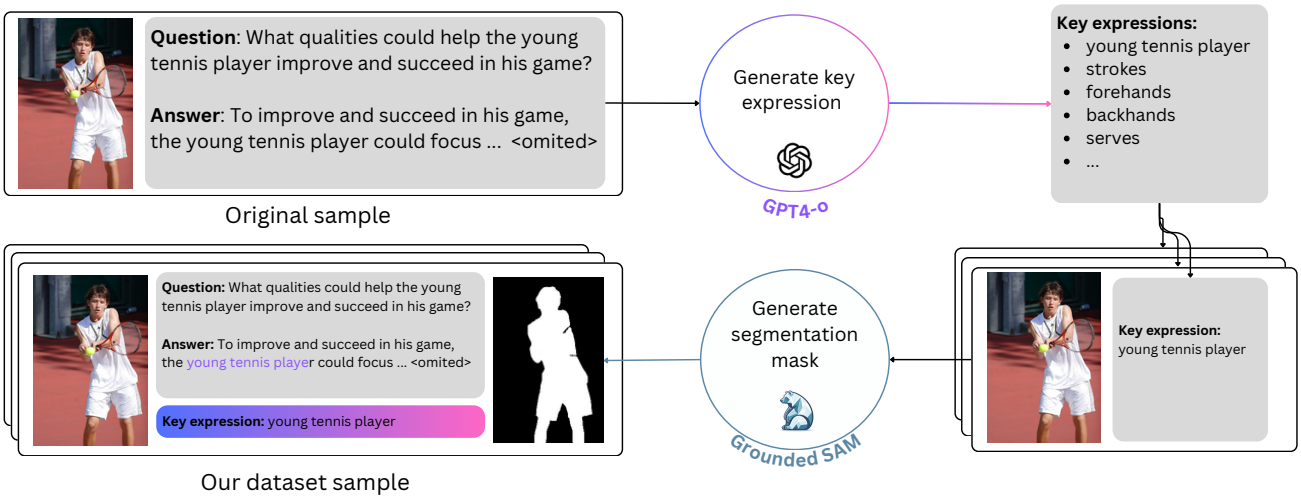


Figure 1. Dataset Collection Overview: The top figure illustrates that GPT is initially used to extract key expressions, while the bottom figure demonstrates that grounded SAM is used to extracts segmentations based on the provided keyword and image.

sion. For such cases, 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. Images are present in 94% of the conversations. Across the entire dataset, we collected around 1.5 million unique segmentation masks, resulting in an average of 2.3 segmentation masks per conversation. This corresponds to about 2.3 million key expressions, averaging 3.5 key expressions per conversation. The range spans from 0 key expression for samples without images or identified keywords to a maximum 33 per conversation. On average, each key expression contains 2.4 words. The average area of the segmentation is 28% of the image.

3.3. Evaluation Datasets

To assess the visual reliance of various LVLMs, we created three benchmark datasets derived from the validation sets of POPE [14], VQAv2 [8], and GQA [10]. These augmented datasets, named FiVL-POPE, FiVL-VQAv2, and FiVL-GQA will be used to assess the extent to which models rely on visual information for accurate responses. We selected theses 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.

Following the procedure outlined in Section 3.1, for each sample, we identified key expressions that can be extracted from the question but also the answer and produce the related segmentation mask. We included the answers as well, as many samples feature the key element in the question, and a significant portion of answers are simple responses like "Yes," "No," or numerical values. Section 5.2 describes our use of these augmented datasets to evaluate how effectively models focus on visual cues. Similar to FiVL-Instruct dataset, GPT-4o does not consistently identify key expressions from each question-answer pair of these three datasets. Moreover, even for samples with key expressions, not all have corresponding segmentation masks generated by the GroundedSAM pipeline. This indicates that some of the questions from these datasets might not depend on the images. In our evaluation set, we filter those cases without extracted keywords or segmentation. Table 1 compares the size of our datasets with the original datasets and Table 2 shows the other statistic of our datasets.

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

Table 1. Sizes of the evaluation datasets after filtering out samples lacking 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 2. Statistics of our evaluation datasets. First row details the average number of key expressions per sample, second row describes the average number of distinct segmentation masks per sample and last row describes the average percentage of the pixels that were masked per sample.

4. Method Evaluation

To ensure the quality of our datasets and effectiveness in training LVLMs for visual grounding, 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 12us different annotators. The accumulated results of the three types of question is that 77% of the annotators labeled the samples as overall good data points. In detail, for the key expression evaluation, 75% of key expressions were marked as essential to the answer's meaning. For the segmentation map evaluation, 58% of segmentation map were annotated as relevant to the key expression. This 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. We also find that the quality of the segmentation is related to the size of the segmentation. 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. Specifically, Figure 2 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. We also made ablations by using FiVL-Instruct filtered to exclude big segmentation masks and compared it with that of training on the unfiltered version and observed no difference in performance. Regardless, this metric can be used as a threshold filtering method for future applications of the dataset. A snapshot of our API is provided in Appendix A.

4.2. Automatic Evaluation

Manual human evaluation is time-consuming. Inspired by recent applications using GPT as an evaluation tool [35], 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 dataset.

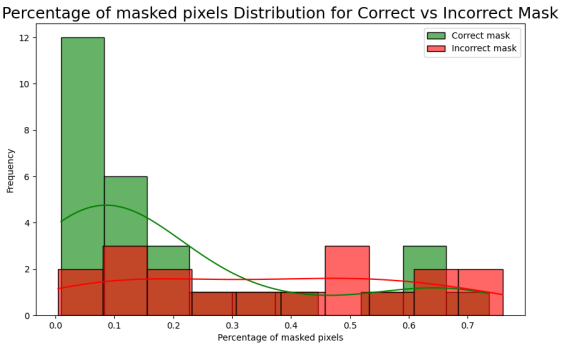


Figure 2. Impact of the size of the segmentation mask. Comparison of the Percentage of masked pixels Distributions for Correctly and Incorrectly annotated Masks

4.2.1. Keyword Evaluation

We prompt the model to verify if the keyword is important to the question and depending on the image. Additionally, we ask the model to rate the degree of importance on a scale from 0 to 10. The full prompt is provided in Figure 4. We report three metrics. The first is the *Importance Ratio = 76%* representing the percentage of extracted expressions classified as key expressions. This result is close to human evaluation, which is 75%. The second is the *Overall Importance Degree = 6.8*, which indicates the average importance score across all keywords, regardless of whether GPT-4o classifies them as important or not important. The third metric is the *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 Figure 4. Results show that only for $Seg1 = 46\%$ of the cases GPT-4o capture the keywords in the segmentation. On one hand, this results 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

```
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 3. Keyword Verification Prompt for GPT-4o.

```
[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 4. Segmentation Verification Prompt for GPT-4o.

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 our training setup, we will take this into account and only focus on the nouns of the key expressions. 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 the fact that 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 Dataset

In this section, we describe three ways 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 novel pretraining task, Vision Modeling. To assess the effectiveness of this task, we fine-tuned an LVLM, specifically, LLaVA-1.5-7b [17], 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. By leveraging this dataset, we ensured alignment with widely used benchmarks in vision-language research, facilitating comparability and reproducibility across studies.

Method. In order to train Large Language Models and even some LVLMs (such as LLaVA), the cross-entropy loss is used to minimize the difference between the predicted probability distribution of the text output and the true distribution of the target words. LLaVA training strategy is in two stages: the first (pretraining) trains a projector which aims to align visual and textual representations, while the second (finetuning) performs only language modeling on the textual outputs of the LM head. In this work, we propose to also guide the visual outputs of the last linear layer (the LM head). Figure 7 plots were created by taking the $argmax$ of the vision logits, mapping it back to the text vocabulary, and highlighting the image patches that relate to that token. We can see from that Figure, column on the right, that the baseline is already able to capture a relevant token from the vocabulary with respect to the image patches. However, since it was not trained to do so, we obtained a lot of different tokens that realize the argmax for each patch, some of them not being relevant. These examples were extracted from samples of LLaVA-1.5-mix-665K instruction dataset. Our dataset gives us access to segmentation maps inherently provided by the vision logits argmax. We augment 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:

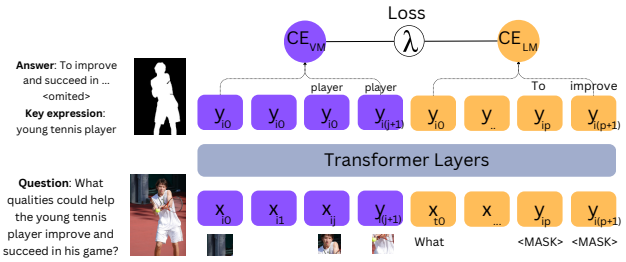


Figure 5. Overview of Vision Modeling pretraining task

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

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. For instance: if given a picture of a bowl of soup the question is *What qualities could help the young tennis player improve and succeed in his game?*. Along with it, our dataset provides the segmentation map related to the *young tennis player*. The LM loss would only guide the relevant tokens y_t to be *To improve and succeed in ... ¡omited¿*. 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: *player* from the text vocabulary. See Figure 5 for a visual explanation of the method. 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). If an object (containing several patches) gets more than one text token assigned to it, we randomly select one. That way, for each image patch, there is maximum one key token that describes the patch. Image patches that do not have a related keytoken are ignored in the loss, similar to LM. We then compute the cross-entropy, CE_{VM} between the created vision labels and the visual logits. We then compute a weighted sum between this loss and the cross-entropy related to language modeling, CE_{LM}. The resulting loss is computed as such:

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

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 [36] 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. (Results were equivalently good for 0.2). Figure 6 shows how we outperformed the baseline in different benchmarks. We outperform the baseline for OK-VQA [24], for which our model achieves 0.55 of accuracy compared to 0.53 for the baseline, MME-cognition [7] for which our model achieves a score of 351 compared to 338 for the baseline, for POPE precision [14] for which our model achieves a precision of 0.95 compared to 0.94 for the baseline, ScienceQA [22], for which our model gets an accuracy of 0.70 compared to 0.66, for MMBench [21] where we obtain 51.6 compared to 50.1, LLaVA-Bench-COCO [17], where we obtain 96 against 93.7 and for LLaVA-in-the-wild [17] for which we obtained 52.6 against 50.3. For Text-VQA [29], VizWiz-VQA [9], GQA [10], POPE accuracy [14] we get comparable accuracies of respectively 0.46, 0.54, 0.62, 0.86. We do notice a small drop in performance for MME-perception [7] from 1470 to 1509.

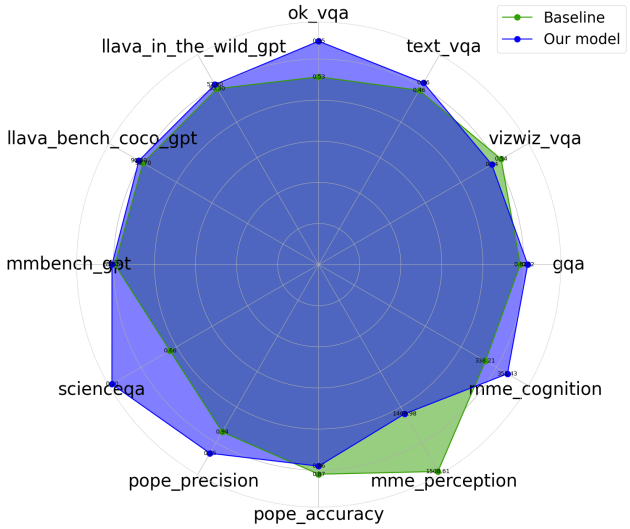
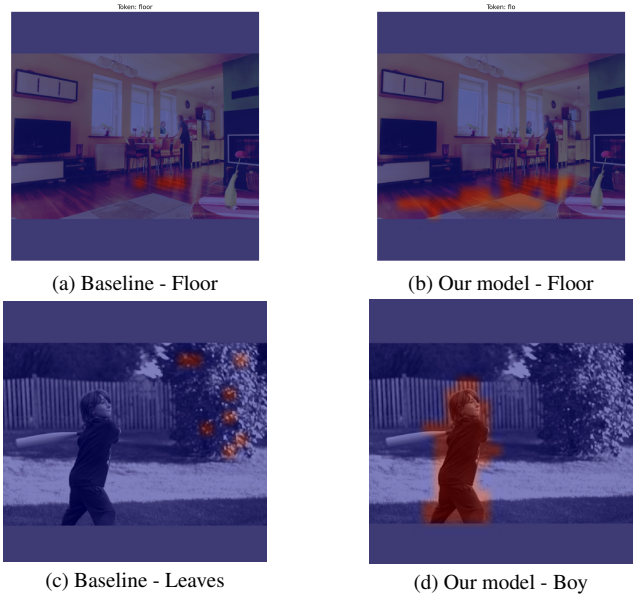


Figure 6. Our model trained on FiVL-Instruct evaluated on various benchmarks compared to the baseline.

As an additional outcome of the training, we also obtain "segmentation maps" by predicting text tokens using an argmax operation over vision logits, which are the outputs at the positions of vision tokens. As a simple observation, averaging over 100 examples, the baseline predicts 74 different tokens overall (with lots of unrelated tokens such



(a) Baseline - Floor (b) Our model - Floor

(c) Baseline - Leaves (d) Our model - Boy

Figure 7. Example of predicted text from vision logits and their corresponding region in the image. The left column displays the vision logits from the baseline model, while the right column shows the output from our model.

as "a", "*", "is" etc.), while our model only encompasses 9 tokens. This demonstrates potential in leveraging visual logits for segmentation, as shown in Figure 7.

5.2. Visual Reliance Evaluation

FiVL datasets also allow us to measure *Visual Reliance* by performing perturbation based evaluation: first evaluating model accuracy on the original, unmasked images, then assessing performance on the masked images. Then, we design a *Visual Reliance Score* (Eq.1), which indicates the percentage of drop in accuracy from the original to the masked image version, with higher scores indicating stronger dependency on visual input. A higher score indicates a greater drop in performance from the original image to the masked image, suggesting a stronger reliance of the model on visual information for generating answers.

$$\text{Visual Reliance Score} = \frac{\text{accuracy}_{\text{original}} - \text{accuracy}_{\text{perturb}}}{\text{accuracy}_{\text{original}}} \quad (1)$$

To validate that FiVL is a good dataset to evaluate visual reliance, we generated a control dataset with random masking. In this control set, each image includes a black rectangular 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 [30], across the three evaluation datasets we created.

Compare FiVL and Random Perturbation. Table 3 shows that across all benchmarks and models, the perturbation based on FiVL masks causes a significantly larger performance drop compared to random perturbation, indicating that our bounding boxes capture meaningful visual content relevant to the questions and FiVL represent good testbeds for visual reliance. One thing to note is that for LLaVA-13B on VQA-v2, we see that masking with random bounding box even improves over the original images. We assume that sometimes random masking may hide non-essential parts of the image, allowing the model to better focus on the remaining regions, which likely contain the intended area of focus.

Comparison LLaVA and Qwen Models. For the VQA-v2 and POPE, we observe that LLaVA-13B shows a greater drop in accuracy compared to Qwen2-VL-7B-Instruct. For GQA, we see the opposite behavior. Since LLaVA-13B exhibits a larger drop on more datasets and the average drop across all three datasets is a bit higher compared to Qwen2-VL (0.513 versus 0.497, respectively), this suggests that the LLaVA-13B model has a higher visual reliance on visual information than the LLaVA model. However, we acknowledge that three datasets may not be sufficient to draw a definitive conclusion, and expanding the evaluation set is a goal for future work.

Compare Three Evaluation Sets. Across models, we observe a larger performance drop on VQA-v2 compared to POPE, and a larger drop on POPE compared to GQA, indicating that VQA-v2 relies most heavily on the image for answering questions, followed by POPE and then GQA.

5.3. Explainability

Here, we also show that FiVL can assist the interpretability of black box models. We produce a summary plot that displays a vision-alignment metric calculated across all heads and layers, like introduced in [2]. Figure 9a offers a detailed view of how the model grounds vision and language through

	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 3. Percentage of drop in performance using the perturbation based method. Using FiVL bounding boxes to introduce perturbations compared to random perturbations across different benchmarks and models.

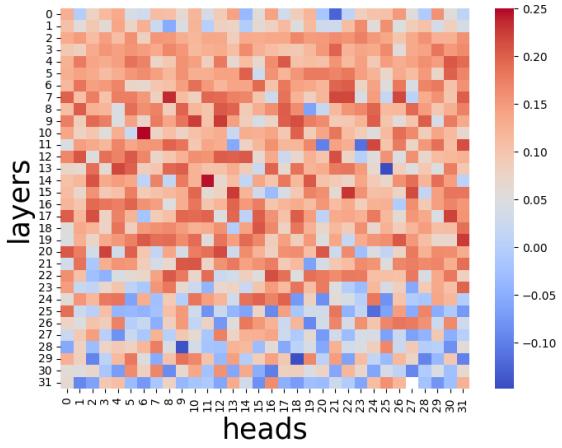
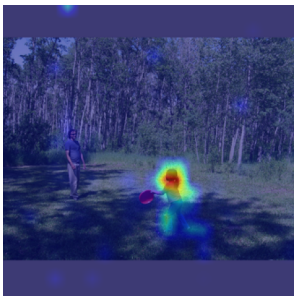


Figure 8. Head summary for VL alignment computed using the Spearman correlation between the segmentation of a token and its related vision attention

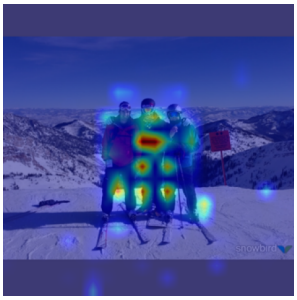
the attentions of each head across all layers. For an attention matrix of size $(N_{layers}, N_{heads}, N_i + N_t, N_i + N_t)$, The head summary calculates the statistical mean over the last two dimensions, producing a plot with dimensions of (N_{layers}, N_{heads}) averaged for 500 samples. For a given question, image, key expression and related segmentation mask from the FiVL-Instruct dataset, we generate the answer using LVaVA-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. Figures 9a and 9b show from which patches of the image the token *girl* gets the most attention, clearly focusing on the girl depicted in the image.

(a) Attention Head (10,6) of the token "girl". Q - *Who are the two people playing Frisbee in the image?* A - *The two people playing Frisbee in the image are a man and a little* ***girl***



(b) Attention Head (14,11) of the token "girl". Q - *Who are the two people playing Frisbee in the image?* A - *The two people playing Frisbee in the image are a man and a little* ***girl***

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



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

Figure 9. 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.

6. Conclusion

In this paper, we introduced FiVL, a framework designed to enhance vision-language alignment and visual focus 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 our FiVL dataset, we saw improvement in a majority of test benchmarks. Our evaluation datasets facilitated comparisons between models regarding their reliance on images to answer questions and provided insight into the degree of image dependency needed across benchmarks. Finally, our explainability application enables users to identify attention heads that excel in vision-language alignment, allowing for a deeper understanding of the internal workings of LVLMs and potential model refinement. Our training method produced a built-in feature that segments the image. Future work could include the evaluation of these segmentations and the enhancement of grounded LVLMs using FiVL.

References

[1] 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. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 12

[2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers, 2022. 7

[3] 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. Pixtral 12b, 2024. 12

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[7] 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. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 6

[8] 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, 2017. 3

[9] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018. 6

[10] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. 3, 6

[11] 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. Segment anything. *arXiv:2304.02643*, 2023. 2

[12] 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. 2

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 12

[14] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 3, 6

[15] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 2

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 5, 6, 12

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[19] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng.

A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[20] 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. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 2

[21] 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. Mmbench: Is your multi-modal model an all-around player?, 2024. 6

[22] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. 6

[23] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024. 1, 2

[24] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. 6

[25] OpenAI. Gpt-4o system card, 2024. 1, 2, 12

[26] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 2

[27] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13009–13018, June 2024. 1, 2

[28] 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. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2

[29] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. 6

[30] 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. 1, 2, 7

[31] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[32] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-Grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 1, 2

[33] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. 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, June 2024. 1, 2

[34] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 2

[35] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 4

[36] 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. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 6

[37] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

Appendix

A. 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 11 and prompts used for evaluation datasets VQA-V2, GQA, and POPE in Figure 12. For each benchmark we use different examples that suit the best to the types of questions. See Figure 13 for FiVL-VQAv2 and Figure 14 for FiVL-GQA and FiVL-POPE.

B. Training details

To train our model, we used Nvidia RTX A6000 GPUs using the hyperparameters from Table 4

Batch Size	4
Number of GPUs	8
Gradient Accumulation	4
Number of epoch	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-patch14-336
Language Model	lmsys/vicuna-7b-v1.5

Table 4. Hyperparameters to train our model

C. 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 5, 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 10 shows the segmentation maps we obtained for the max-v token describing each image. For example for the example 10a, 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 5. Performance of the segmentation maps inherently provided by our model

D. Additional Visual Reliance evaluations

To expand our benchmarks and gain a broader understanding of model/benchmark performance, we evaluated five additional models on FiVL-VQAv2, FiVL-POPE, and FiVL-GQA. This helps to assess the generalizability of our approach across more models. Table 6 shows our results for

(a) Bear

(b) Bird

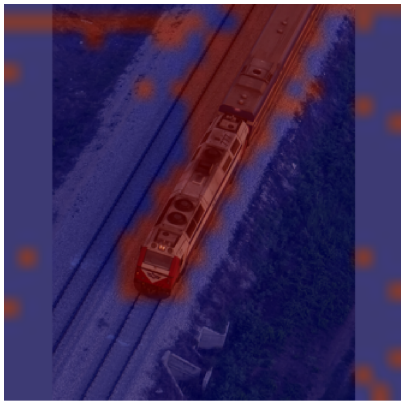
(c) Birds

(d) Bott

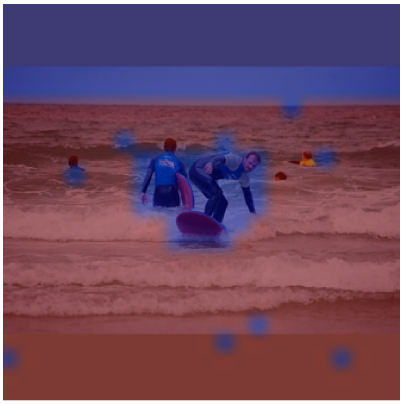
(e) Chair

(f) Dog

(g) People



(h) Train



(i) Water

Figure 10. 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

the 2 models presented in Section 5.2 as well as for LLaVA-v1.5-7b [17], GPT4o [25], BLIP-2 [13], Pixtral-12B [3] and Phi3-Vision[1], which are state-of-the arts mutlimodal models. In bold, are the highest visual reliance scores per model and across all benchmarks. Underlined are the highest visual reliance scores across models, given a benchmark.

CVPR
#12778

The results unanimously indicate that, among all models, VQA-v2 appears to be the benchmark that relies most heavily on the image to answer the questions. Looking at the average performance per model across benchmarks (last column), we observe that GPT4o relies most heavily on the image as a reference for answering, followed by Pixtral-12B. We can also observe that BLIP-2 is minimally affected by the adversarial attack indicating that it does not heavily rely on the image to generate answers. This lack of reliance could explain its low accuracy on the original images for this benchmark (30 % compared to approximatively 80% for the other models). As a result, perturbing the images had little impact on its performance (29 %).

	VQA-v2	GQA	POPE	Average
Qwen2-VL-7B	**0.64**	0.39	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
Average	**0.65**	0.40	0.42	-

Table 6. Visual reliance scores: % of drop in performance using FiVL bounding boxes for perturbations. In **bold**: highest scores across benchmarks. Underlined: highest scores across models

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 11. FiVL-Instruct system prompt used for training datasets

CVPR
#12778

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 12. 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 13. Examples for FiVL-VQAv2

CVPR
#12778

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 14. Examples for GQA and POPE prompts

E. API of the manual evaluation

Figure 15 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?

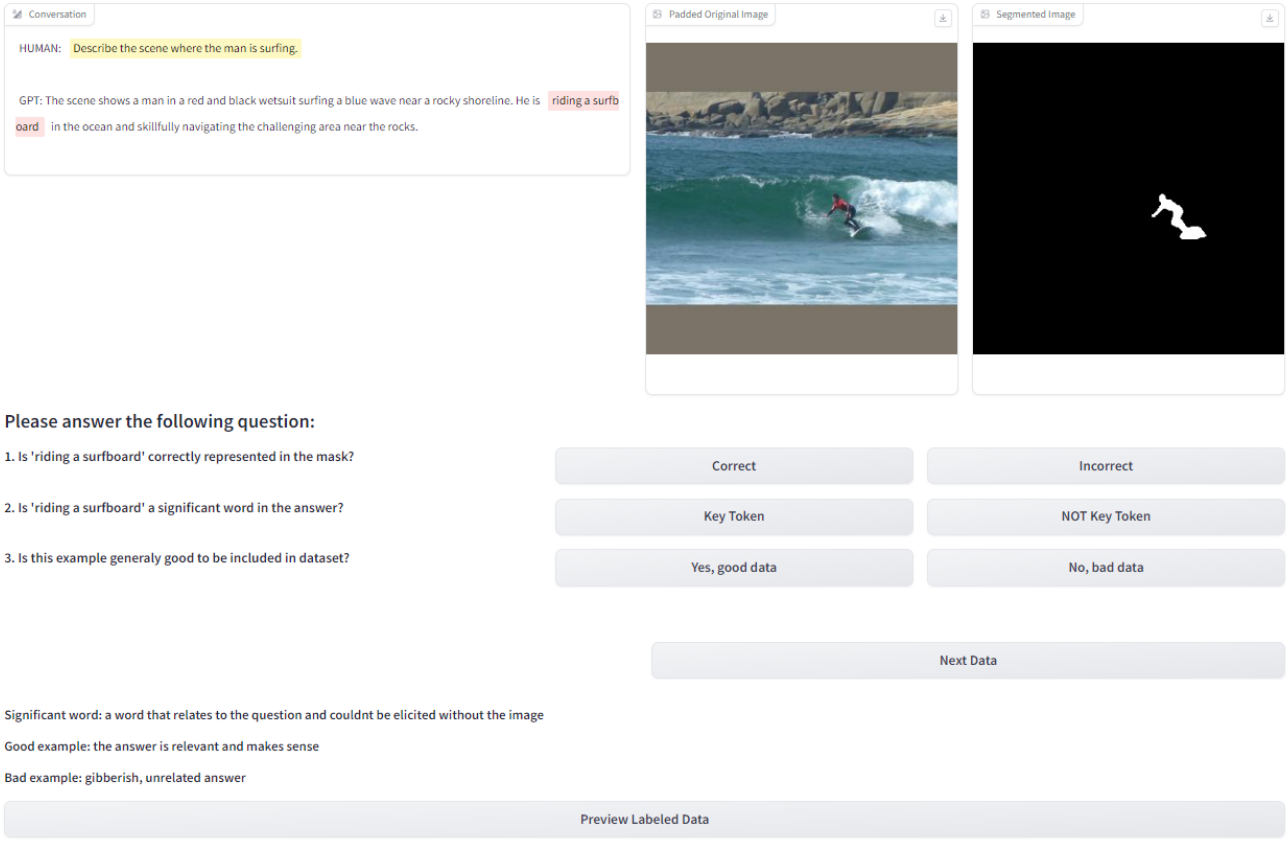


Figure 15. Web user interface for our dataset evaluation