LANGUAGE-INSTRUCTED VISION EMBEDDINGS FOR CONTROLLABLE AND GENERALIZABLE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models are typically trained as static feature extractors, forcing the burden of task adaptation onto large downstream models. We propose a different paradigm: instead of solely feeding visual features into language, we use language itself to dynamically guide the vision encoder. Our method, Language-Instructed Vision Embeddings (LIVE), leverages language as high-level guidance to produce task-centric embeddings at inference time—without requiring task-specific retraining. This enables the encoder to focus attention on contextually relevant aspects of the input, yielding more controllable and generalizable representations. Empirically, LIVE reduces visual hallucinations (+34 points on MMVP), outperforms vision–language models with orders of magnitude more parameters on visual question answering, and generalizes to unseen instructions and tasks—offering a direct path toward adaptive, instruction-driven visual intelligence.

1 INTRODUCTION

A hallmark of human vision is its active, selective nature. Guided by internal goals or task demands, we focus on relevant parts of the visual world while ignoring distractions (Posner, 1980; Desimone et al., 1995). When searching for a specific object or understanding a particular interaction, humans implicitly "know" where and what to look for. In contrast, today's leading vision models, despite producing powerful general-purpose features (Oquab et al., 2023; Zhai et al., 2023; Yu et al., 2022; Radford et al., 2021), lack this dynamic, intention-driven adaptability. Their representations are typically *static*, pre-computed without reference to the specific query they are meant to serve.

This limitation is particularly acute in vision–language models. Existing approaches such as visual prompting (Bahng et al., 2022; Shtedritski et al., 2023) or fine-tuning (Mao et al., 2022) provide limited adaptability, but they cannot interpret zero-shot language instructions. Dominant LLM-centric architectures (Liu et al., 2023; Grattafiori et al., 2024; Alayrac et al., 2022; Team, 2024) delegate language integration to large downstream modules, incurring high computational cost while being unable to recover fine-grained details overlooked by the vision encoder, often resulting in hallucinations (Tong et al., 2024). Recent attempts to modulate vision encoders with paired captions (Lavoie et al., 2024; Xiao et al., 2025) are restricted by their reliance on descriptive text rather than true instructions, limiting their versatility and controllability. Thus, the central challenge remains: how to embed language-driven control into the vision encoder to yield adaptive, task-aware representations.

We address this challenge with LIVE (Language-Instructed Vision Embeddings), a simple and effective framework for creating **language-steered vision embeddings**. LIVE enables dynamic, fine-grained control of a vision encoder by training it to follow textual instructions. Concretely, we use a large language model (LLM) as the knowledge base to generate synthetic instruction–response pairs, which we combine with images into contrastive triplets. This teaches the vision encoder to steer its embeddings based on textual commands, allowing it to highlight relevant attributes or suppress adversarial cues (e.g., typographical attacks), thereby achieving robust instruction-following at the representation level (Figure 1).

Once trained, LIVE yields standalone, language-steered embeddings that downstream tasks can use directly—no large LLMs or task-specific fine-tuning required. Trained on synthetic ImageNet-based data, LIVE generalizes strongly to real, unseen tasks: it reduces hallucinations by 34 points on MMVP (Tong et al., 2024), and surpasses LLM-based counterparts on GQA (Hudson & Manning, 2019) by 7 points with less than 10% of their parameters. We also measure and narrow its gap to its

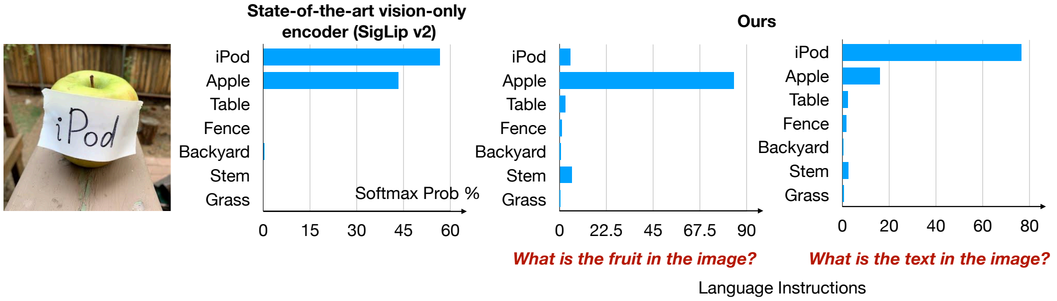


Figure 1: **LIVE (Language-Instructed Vision Embedding).** We show state-of-the-art vision foundation model struggle to distinguish between text and objects. LIVE allows user-guided focus on specified aspects (*e.g.*, "fruit" *v.s.* "text"), boosting control and prediction accuracy.

LLM knowledge base by up to 49 points across instruction-following benchmarks. Attention and retrieval visualizations show precise, instruction-driven control emerging inside the encoder. With up to 10 times fewer parameters than LLM-heavy methods, our results suggest that embedding task instructions into the vision encoder—rather than scaling downstream modules—is an efficient path to generalizable, and controllable visual perception.

2 Related Work

Vision Foundation Models. Recent vision foundation models often use two-tower architectures and train contrastively with image-text pairs (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025; Zhai et al., 2022b). While some approaches jointly optimize contrastive and generative objectives (e.g., CoCa (Yu et al., 2022), Mammut (Kuo et al., 2023)) or use encoder-based captioning (e.g., Flamingo (Alayrac et al., 2022), Pali (Chen et al., 2022)), the vision embeddings are typically computed independently or language interaction occurs during late fusion transformers. Similarly, methods like Q-former (BLIP-2) (Li et al., 2022) use intermediate stages and powerful LLM decoders for image-to-text tasks, without directly instructing the frozen image encoder with language.

Alternative paradigms like masked image-text modeling (ViLT (Kim et al., 2021)) learn alignments but are not optimized for retrieval embeddings, therefore they require further finetuning on the target task and cannot perform prediction in a zero-shot manner. More recent architectures aim to unify approaches (X-Former (Swetha et al., 2024)) or leverage LLMs as decoders for richer outputs and supervision (Lin et al., 2021; Liu et al., 2023; Beyer et al., 2024; Team et al., 2025; Grattafiori et al., 2024; Wan et al., 2024; Tschannen et al., 2023; Lin et al., 2024). However, these methods still generally do not allow language to directly control the vision embeddings and cannot perform the target task via retrieval. Alternative vision only models, like Dino (Oquab et al., 2023; Caron et al., 2021) and JEPA (Assran et al., 2023), cannot handle language inputs. BRAVE (Kar et al., 2024) ensembles vision encoders for improved accuracy.

Instructed Foundation Models. The growing need for adaptive vision-language models inspired efforts in fine-tuning (Lin et al., 2023) and prompt engineering (Menon et al., 2022). However, these approaches typically optimize either the entire model or specific prompts, restricting them to single-task adaptations such as rationale explanation (Mao et al., 2023) or category classification (Mao et al., 2022). Methods like (Shtedritski et al., 2023; Zhong et al., 2022) allow querying visual encoders via explicit markers (e.g., red circles or bounding boxes) but fail in scenarios involving overlapping or ambiguous visual concepts, as these markers only specify location without clarifying the targeted attribute (e.g., color, texture). Prior work train top down vision encoder for embodied agent, yet this is not zero-shot (Eftekhar et al., 2023). Magiclens (Zhang et al., 2024) perform self-supervised image retrieval based on instructions, yet it does not provide retrieval in semantic, language space, and not ready for direct visual perception.

Multimodal retrieval methods like UniIR (Wei et al., 2024) perform retrieval via late-stage fusion of features, our work focuses on guiding the vision encoder itself, which could serve as an enhanced vision component to complement such models. (Kar et al., 2024) combines multiple vision encoders to obtain better vision representations for language models. Other approaches control vision indirectly through post-hoc modification (Chen et al., 2024a) or in specialized domains like document retrieval (Zhou et al., 2024; Chen et al., 2024b). A recent trend is fine-tuning vision LLMs for

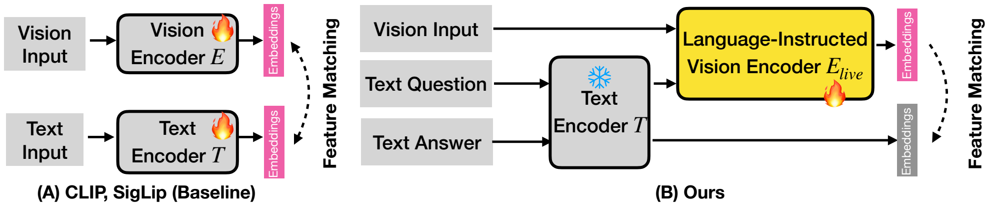


Figure 2: **Instructive Vision Encoder Design.** Prior vision-language models like CLIP (Radford et al., 2021) and SigLip (Zhai et al., 2023) use two-tower architecture with separate vision and text encoders. We reuse the text tower to embed the query, apply a projection layer, and feed it into the vision transformer alongside with the input image. Note that for the text question and the text answer, while they are feed into the same text encoder, they are processed separately and there is no feature interaction. The text encoder is frozen; only the vision encoder is trained (yellow). We denote learnable embeddings in pink.

retrieval (Wei et al., 2024; Jiang et al., 2024; Liu et al., 2025). While powerful, these models inherit the substantial computational footprint of their underlying LLMs. The most related works that also modulate the vision encoder directly typically use image captions as the conditioning signal (Lavoie et al., 2024; Xiao et al., 2025). This strategy risks learning undesirable shortcuts, as the model can minimize loss by simply matching text features rather than learning true visual grounding. Our method, LIVE, explicitly decouples the guidance from the target by using task instructions that differ significantly from the target description. This design forces the model to learn a more sophisticated mechanism for instruction-based control, enabling precise manipulation of vision embeddings without the inference overhead of large LLMs or the risk of learning trivial solutions.

3 Methodology: Learning Language-Instructed Vision Embeddings

Conventional vision-language models treat the vision encoder as a static feature extractor for a downstream LLM. We invert this paradigm by distilling knowledge from an LLM back into the vision encoder itself. Using synthetic image-query-answer triplets, we employ contrastive learning to train the encoder to produce language-instructed embeddings that align with the answer's semantics. We will show a powerful, standalone vision encoder capable of zero-shot perception, eliminating the need for a computationally expensive LLM at inference.

3.1 Language-Instructed Visual Embeddings

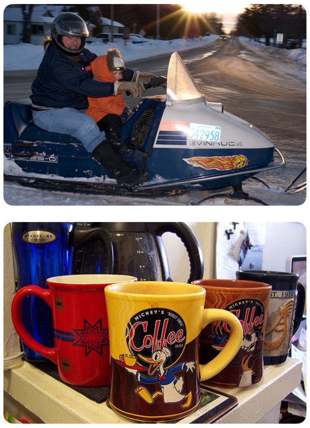
Standard vision-language models like CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) use a two-tower architecture with vision $E(\cdot)$ and text encoders $T(\cdot)$, see Figure 2. The vision encoder E generates a general-purpose embedding $\mathbf{z} = E(\mathbf{x})$ intended to capture all relevant information in the input image $\mathbf{x}$. However, to serve diverse downstream tasks via text queries, these representations must be versatile and precise. Training such universal embeddings is challenging due to vision encoder capacity limits.

Visual prompting (Bahng et al., 2022; Jia et al., 2022) aims to adapt visual representations. However, visual perception is often ambiguous and context-dependent. Simple location prompts (e.g., boxes, circles (Shtedritski et al., 2023)) offer some control but lack granularity to specify which aspects. For example, a car region might require focusing on color (”What color?”), make (”What make?”), or condition (”Is it clean?”). Existing methods struggle with this semantic ambiguity.

To address this, we propose an language-conditioned vision encoder, denoted E_{live}. Instead of a fixed embedding, E_{live} dynamically processes the image $\mathbf{x}$ based on the embedding of a textual instruction $\mathbf{q}$. We reuse the pretrained text encoder $T(\cdot)$. Our instructive visual embedding is computed as:

$$\mathbf{z}^{(I)} = E_{live}(\mathbf{x}, T(\mathbf{q})). \tag{1}$$

This formulation allows the vision encoder to focus its computation on the aspects of the image most relevant to the language instruction, producing a targeted, task-specific representation. Model implementation details are in Section 4.1.



Question	Answer
What is the primary activity taking place in the image?	A person is driving a snowmobile with a child passenger.
What type of vehicle is depicted in the image?	A vintage snowmobile.
What is the color of the snowmobile?	The snowmobile is predominantly blue and white.
What is the color of the child's jacket?	Orange
What is the adult wearing on their head?	A helmet.
What is visible in the background of the image?	Houses, trees, a road, and a bright sun
What is the license plate number on the snowmobile?	542958
What is the color of the mug in the center of the image?	The mug in the center is primarily yellow and brown.
What character is featured on the yellow mug?	The yellow mug features Donald Duck.
What object is behind the mugs?	A coffee maker is behind the mugs.
What is the relative position of the red mug to the yellow?	The red mug is to the left of the yellow mug.
Is there coffee in any of the mugs?	Yes, there appears to be coffee in the yellow mug.
What kind of surface are the mugs sitting on?	The mugs are sitting on a white shelf or counter.
If the foreground is ignored, what is the main object in the image?	If the foreground is ignored, the main object in the image is the coffee maker

Figure 3: **Triplet Training Data from LLM.** We apply Gemini-2.0-Flash (Comanici et al., 2025) to automatically create diversified, open-world triplet data containing image, query, and answer. This method moves beyond generic questions from existing image-text data, allowing for nuanced and sophisticated exploration of individual image content.

3.2 Training Objective

We train the instruction-conditioned vision encoder E_{live} by matching its output $\mathbf{z}^{(I)}$ with the text embedding of the corresponding correct answer $\mathbf{a}$. Specifically, we want $\mathbf{z}_i^{(I)} = E_{live}(\mathbf{x}_i, T(\mathbf{q}_i)$ to be close to the answer embedding $\mathbf{z}_j^{(T)} = T(\mathbf{a}_j)$ if and only if $(\mathbf{x}_i, \mathbf{q}_i)$ corresponds to answer $\mathbf{a}_j$.

Following (Zhai et al., 2023), we employ a sigmoid-based alignment loss, which yields better performance and stability than standard contrastive losses (Radford et al., 2021). Given a batch of image-instruction pairs $(\mathbf{x}_i, \mathbf{q}_i)$ and their corresponding answers $\mathbf{a}_j$, the loss is defined as:

$$\mathcal{L} = -\mathbb{E}_{i,j}\left[\log \frac{1}{1+\exp(-y_{ij}(t(\mathbf{z}_i^{(I)} \cdot \mathbf{z}_j^{(T)}) + b))}\right] \tag{2}$$

$y_{ij} \in \{-1, 1\}$ encodes the match (1 for match, -1 for mismatch). The parameters t (temperature) and b (bias) are learnable parameters for calibration. We minimize this loss via gradient descent to optimize the visual encoder E_{live}.

3.3 Knowledge Transfer from LLM

Despite the abundant image text paired data (Schuhmann et al., 2022; Byeon et al., 2022), a significant challenge in training the instruction-guided encoder E_{live} is the scarcity of large-scale datasets with image-instruction-answer triplets $(\mathbf{x}, \mathbf{q}, \mathbf{a})$. Existing visual question answering (VQA) datasets (e.g., CC3M-VQA (Changpinyo et al., 2022)) often rely on template-based or rule-generated questions, which may not capture the breadth and complexity of real-world queries needed to probe deeper understanding. Our experiments in Figure 7 shows existing datasets cannot train language-instructed visual embeddings with high accuracy.

To overcome this data bottleneck, we leverage the extensive world knowledge and reasoning capabilities inherent in LLMs. We treat an LLM as an implicit knowledge source capable of identifying salient aspects of an image and formulating relevant questions about them. We query powerful LLMs that take visual inputs, and generate question-answer pairs $(\mathbf{q}, \mathbf{a})$ conditioned on image. This transfers the complex understanding capabilities of the LLM from many billions of parameters into the training data for our vision encoder. Crucially, this computationally intensive LLM inference occurs offline during dataset creation. At inference, our approach only require the efficient instruction-guided vision encoder E_{live}, preserving computational efficiency for real-time perception tasks.

We prompt LLM to generate multiple diverse question-answer pairs for each image simultaneously with the following prompt structure: *Provide a numbered list of interesting visual questions about the image, followed by the corresponding answers.* Figure 3 shows examples of our generated queries

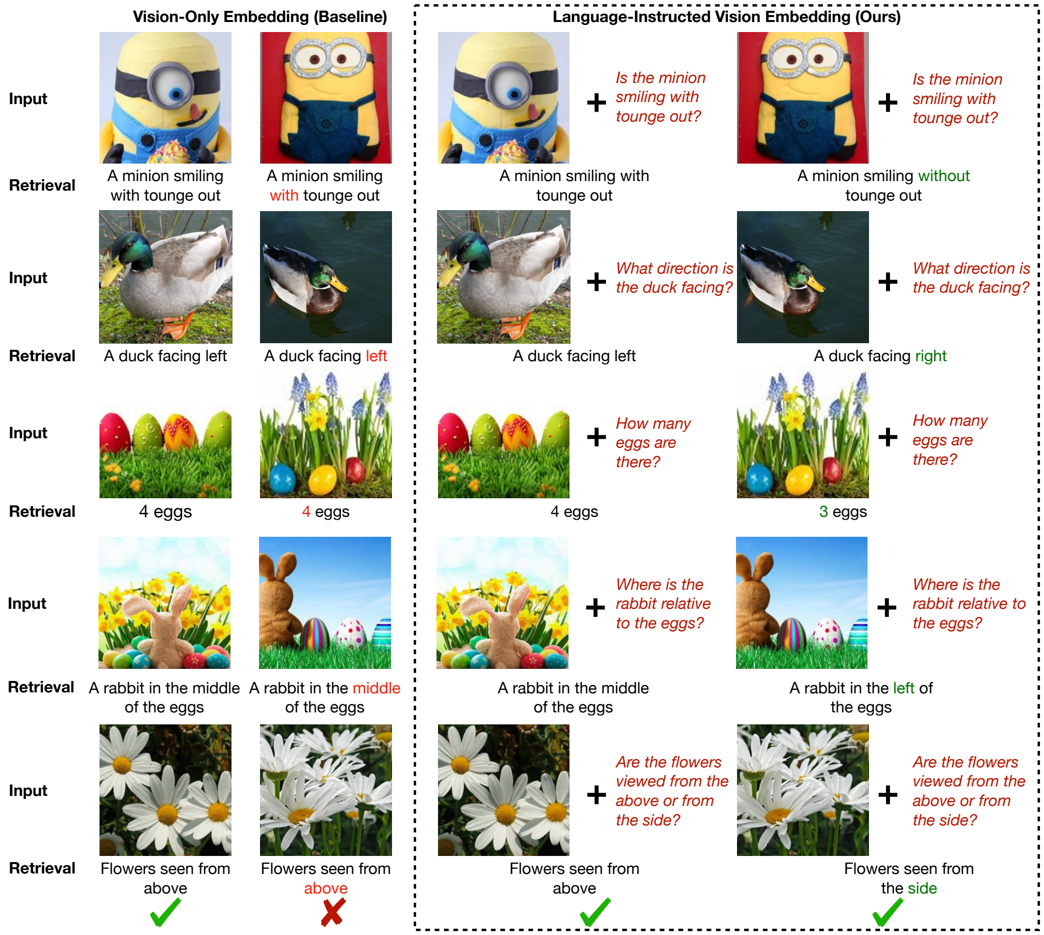

Vision-Only Embedding (Baseline)
Language-Instructed Vision Embedding (Ours)
Input
Is the minion smiling with tounge out?
Is the minion smiling with tounge out?
Retrieval
A minion smiling with tounge out
A minion smiling with tounge out
A minion smiling with tounge out
A minion smiling without tounge out
Input
What direction is the duck facing?
What direction is the duck facing?
Retrieval
A duck facing left
A duck facing left
A duck facing left
A duck facing right
Input
How many eggs are there?
How many eggs are there?
Retrieval
4 eggs
4 eggs
4 eggs
3 eggs
Input
Where is the rabbit relative to the eggs?
Where is the rabbit relative to the eggs?
Retrieval
A rabbit in the middle of the eggs
A rabbit in the middle of the eggs
A rabbit in the middle of the eggs
A rabbit in the left of the eggs
Input
Are the flowers viewed from the above or from the side?
Are the flowers viewed from the above or from the side?
Retrieval
Flowers seen from above
Flowers seen from above
Flowers seen from above
Flowers seen from the side
✓
✗
✓
✓

Figure 4: **LIVE Reduces Visual Hallucinations (MMVP Benchmark Tong et al. (2024)).** State-of-the-art vision-only embeddings Zhai et al. (2023) (left column) must perceive the entire scene without query-specific guidance, making them prone to hallucination when precise details are needed. By modulating visual computation with the input text query (right column), our method selectively focuses on relevant information, thereby mitigating hallucinations and improving accuracy.

and answers on ImageNet, which introduce diverse visual attributes and semantic concepts previously unavailable. This rich, detailed data enables vision models to learn beyond prevalent image-captioning patterns, fostering more effective, fine-grained visual comprehension. Crucially, as humans often perform such queried visual tasks spontaneously using System-1 (intuitive) reasoning, our LIVE encoder is designed to be similarly efficient and sufficient. This avoids the computational overhead of larger LLMs, especially on direct visual perception applications. Moreover, if the downstream application is known before deployment, our text embeddings can be pre-computed and cached, which saves significant computation.

4 Experiment

This section details our experimental setup, benchmarks, baselines, results, and analysis designed to evaluate the zero-shot language controllability enabled by our LIVE approach.

4.1 Experimental Setup

Training Data: We train LIVE using ImageNet training images. We apply the data generation detailed in Section 3.3 on ImageNet data via the public available Gemini 2.0 Flash (Gemini, 2024; Comanici et al., 2025), where we generate a total around 16.4 million images-query-answer triplet

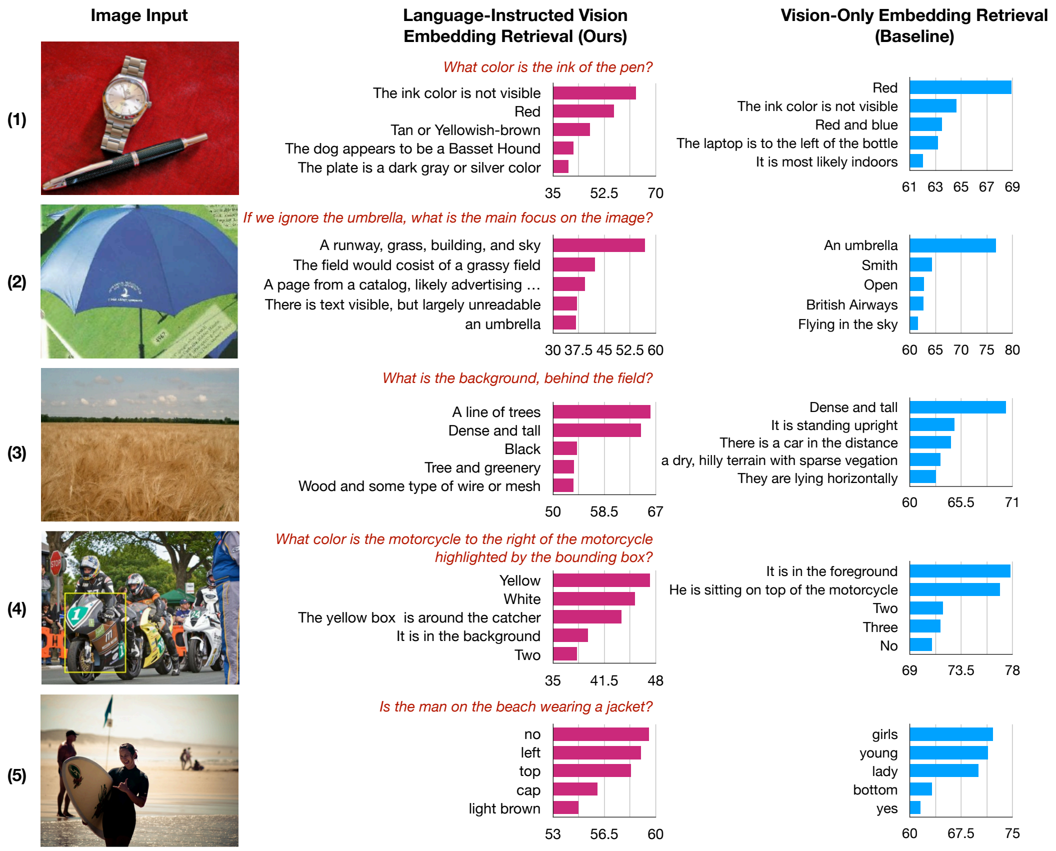


Figure 5: **LIVE's Retrieval based on Language Instructions.** Examples 1-5 show examples from ImageNet, Caltech, SUN, RefCOCO, and GQA, respectively. Instructions provided to the model during inference are unseen in training and in *red*. For both our method and the vision-only baseline (SigLip), we show the top 5 retrieved text responses with bars indicating the predicted sigmoid probability. Our method demonstrates superior retrieval accuracy, correctly identifies (1) non-visible elements, (2) follow instructions to ignore, (3) attend to factors asked, (4) do basic spatial reasoning, (5) perceive relationships.

data since there will be multiple informative queries for each image. In addition, we explore and compare with available triplet data from PaliGemma (Beyer et al., 2024), specifically leveraging CC3M-VQA (Changpinyo et al., 2022). We also create a universal instruction, "caption the image", for dataset that does not have instructions available like WebLI (Wang et al., 2025) and Open Images (Piergiovanni et al., 2022).

Evaluation Benchmarks: We target tasks that require explicit language instructions to specify the goal, in contrast to popular benchmarks that use static vision encoders for universal zero-shot classification and thus cannot probe instruction-following or capabilities beyond fixed label taxonomies. We deduplicate both the image and text instructions to make sure all our evaluations data are novel. For all datasets, we report the Top 1 retrieval accuracy. MMVP (Tong et al., 2024) is a recent popular benchmark that evaluate vision language models' hallucinations. See example in Figure 4. The images are paired so that the model has to see the nuances to answer the question in the right way. GQA (Hudson & Manning, 2019) is a challenging question answering benchmark that goes beyond attribute answering and require reasoning to answer the scene graphs related questions.

Moreover, to measure the gap of LIVE to its LLM knowledge source, we use Gemini 2.0 Flash, to annotate instruction answers on test data on Caltech101 (Griffin et al., 2007), SUN397 (Xiao et al., 2010), RefCOCO (Kazemzadeh et al., 2014), and ImageNet (Deng et al., 2009). Therefore Gemini 2.0 Flash has 100% accuracy. We filtered the test and ensured no instruction overlap between training and test sets. We denote the repurposed dataset with †. Those † datasets's goal is to measure the

	Image Size	Params (M)	🧭	🔍	🔄	⇅	📍	🎨	⚙	A	📷	Average
OpenAI ViT-L-14 (Radford et al., 2021)	224^2	427.6	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 (Radford et al., 2021)	336^2	427.9	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
DFN ViT-H-14 (Fang et al., 2023)	224^2	986.1	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 (Fang et al., 2023)	378^2	986.7	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 (Xu et al., 2023)	224^2	427.6	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 (Xu et al., 2023)	224^2	986.1	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 (Sun et al., 2023)	224^2	1136.4	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ (Sun et al., 2023)	224^2	5044.9	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3
SigLIP ViT-SO-14 (Zhai et al., 2023)	224^2	877.4	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 (Zhai et al., 2023)	384^2	878.0	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
InstructBLIP Li et al. (2022)	336^2	∼14200.0	-	-	-	-	-	-	-	-	-	16.7
LLaVa Liu et al. (2023)	336^2	∼13000.0	-	-	-	-	-	-	-	-	-	31.3
BRAVE Kar et al. (2024)	336^2	∼10300.0	-	-	-	-	-	-	-	-	-	42.0
SigLIP ViT-SO-14 (Ours)	384^2	891.0	**80.0**	**76.7**	**73.3**	**80.0**	**83.3**	**86.7**	**66.7**	**66.7**	**73.3**	**76.3**

Table 1: **Zero-Shot Accuracy on MMVP-VLM benchmark (Tong et al., 2024).** We use **bold** to highlight the best accuracy. Baseline methods are vision-only models, numbers are quoted from MMVP. We symbol the visual patterns following MMVP: 🧭: Orientation and Direction, 🔍: Presence of Specific Features, 🔄: State and Condition, ⇅: Quantity and Count, 📍: Positional and Relational Context, 🎨: Color and Appearance, ⚙: Structural and Physical Characteristics, **A**: Texts, 📷: Viewpoint and Perspective. All CLIP-based methods, using vision-only embeddings, struggle on this benchmarks. Just by instructing the vision embeddings, our method achieves a **34-point** zero-shot accuracy improvement over prior SOTA methods merely, which underscores the importance of our instructions in guiding the vision encoder to focus on relevant signals and reducing hallucinations.

(a) Performance of our model variants on GQA.

ViT Model	SigLip	Fusion	Menon et al.	Ours
SigLip ViT-T-14	9.8	20.0	10.4	**60.6**
SigLip ViT-B-16	12.0	12.8	13.0	**71.2**
SigLip 2 ViT-B-16	14.4	20.8	19.8	**67.6**
SigLIP 2 ViT-SO-14	16.4	19.6	17.8	**68.2**

(b) Comparison with state-of-the-art.

Model	Accuracy (%)
BLIP-2	44.7
InstructBLIP	49.5
BRAVE	52.7
LLava	63.3
Ours (ViT-B-16)	**71.2**

Table 2: **Zero-Shot Retrieval Accuracy on GQA tasks.** We report top-1 accuracy (%).

fidelity of knowledge transfer from the ”teacher” (Gemini) to the ”student” (LIVE), not to benchmark the accuracy on target tasks, since those tasks is Gemini generated.

Baselines. One family of baselines are static vision-only embeddings and their variants. *SigLip* (Tschannen et al., 2025): We use downloaded SigLip models up to size SO400M, representing state-of-the-art static, instruction-agnostic visual embeddings. *Fusion*: We directly add the embeddings of the image and text query, and the goal is to retrieve the text answer. *Menon et al.* (Menon & Vondrick, 2022): Following Menon et al. (Menon & Vondrick, 2022) that achieved improved retrieval accuracy by adding language descriptions to the text answer, we will append the language instructions to the answer as the description for the task, which will tell the model's text tower the query asked. Moreover, We compare with LLM-based approach *LLaVa (Liu et al., 2023)* and late fusion based method *InstructBLIP (Li et al., 2022)*. We also compare with an ensemble-based state-of-the-art method *BRAVE (Kar et al., 2024)*, that combines five generic vision encoders EVA-CLIP-g (Sun et al., 2023), CLIP-L/14 (Radford et al., 2021), SILC-G/16 (Naeem et al., 2024), ViT-e (Chen et al., 2022), and DINOv2-L/14 (Oquab et al., 2023) and further finetune.

Implementation Details. We initialize LIVE's vision encoder from a pretrained SigLip and SigLip-v2 (Zhai et al., 2023; Tschannen et al., 2025), since it outperforms CLIP (Radford et al., 2021). All models are transformer (Vaswani et al., 2017). We use SigLip text encoder to precompute fixed text embeddings for instructions during training and evaluation. The precomputed instruction text embedding is projected by a single linear layer which will be input to the vision transformer, which introduces additionally 13M parameters for ViT-So model. The vision encoder, including the text projection layer, will be updated during training, while the original text tower remains fixed. We use the same optimizer as SigLip (Zhai et al., 2022a) with a learning rate of 0.001, batch size of 8192,

	ImageNet†				Caltech 101†			
ViT Model	SigLip	Fusion	Menon et al.	Ours	SigLip	Fusion	Menon et al.	Ours
SigLip ViT-T-14	25.10	32.46	33.42	**73.28**	10.53	11.38	22.05	**37.08**
SigLip ViT-B-16	30.84	33.23	42.50	**86.93**	12.08	12.64	24.72	**55.75**
SigLip 2 ViT-B-16	37.73	40.69	60.52	**86.79**	14.89	15.31	29.92	**51.97**
SigLIP2 ViT-SO-14	38.03	40.40	60.86	**87.06**	14.61	15.87	33.00	**55.05**

	SUN†				RefCOCO†			
ViT Model	SigLip	Fusion	Menon et al.	Ours	SigLip	Fusion	Menon et al.	Ours
SigLip ViT-T-14	6.99	8.87	10.90	**33.16**	8.52	11.74	11.01	**42.73**
SigLip ViT-B-16	9.26	9.75	16.94	**49.83**	9.84	10.87	12.78	**59.32**
SigLip 2 ViT-B-16	12.44	12.96	24.67	**49.76**	12.04	13.51	17.47	**55.95**
SigLIP 2 ViT-SO-14	13.00	14.06	25.79	**52.94**	9.40	10.28	14.98	**54.33**

Table 3: **Closing the gap to the Gemini knowledge source in zero-shot instruction following.** We report top-1 retrieval accuracy on benchmarks† where Gemini's annotations serve as a 100% accurate oracle. Our model is evaluated in a strict zero-shot setting—without any fine-tuning on the downstream Caltech 101, SUN, or RefCOCO datasets, unlike prior work (Beyer et al., 2024; Kim et al., 2021). All evaluation sets are deduplicated from our synthetic training data. Our approach narrows the gap to the oracle, outperforming baselines by up to 49 points.

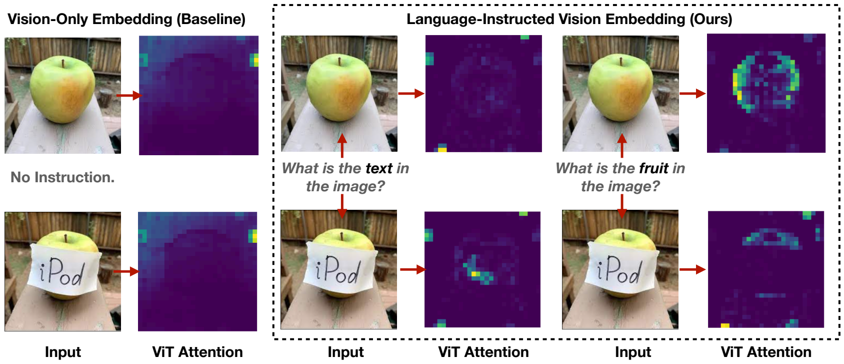


Figure 6: **Zero-Shot Language Instructions Steer Visual Attention.** Unlike baseline encoders producing global attention (SigLip, left), our LIVE uses instructions to focus dynamically. Prompting for "text" highlights the "iPod" label; prompting for "fruit" highlights only the apple, ignoring the label. This demonstrates emergent, instruction-driven control over visual encoding.

and train for 122k steps. We use 256 TPUv3 for training. Following SigLip, we apply only resize augmentation during training.

4.2 Results

We first evaluated LIVE on the MMVP-VLM benchmark (Tong et al., 2024). As shown in Table 1, our model achieves a 34-point accuracy gain over prior established methods, including LLM and ensemble counterparts that are 10 times more larger. Qualitative examples in Figure 4 illustrate how language instructions can guide the vision model to focus on end-task requirements, thereby mitigating hallucinations, such as the erroneous perception of a minion's tongue. On question answering that requires relationship reasoning GQA dataset, we outperform LLMs and prior best generic vision models, by 7 points, with 10 times fewer parameters.

We then assessed our LIVE's acucracy gap to its LLM knowledge source, Gemini 2.0 Flash. Results in Table 3 demonstrate that LIVE, while still has 23 point to 41 points gap from Gemini Oracle, attains superior top-1 retrieval accuracy compared to established vision-only embeddings on these targeted tasks, despite considerable domain shifts in both images and query types.

Figure 5 shows visualizations of images, queries, and top-5 retrieved instances. All are deduplicated from training set. Our LIVE model exhibits emergent capabilities not present in its training. For instance, as illustrated by image (4) in Figure 5, the model correctly interprets bounding boxes to infer the color of a specified motorcycle, despite no exposure to bounding box annotations during training. Furthermore, Figure 5 image (1) demonstrates the model's ability to discern nuanced visual details,

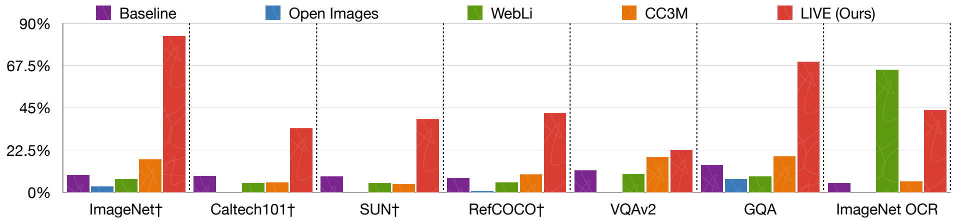


Figure 7: **Impact of Triplet Training Data on LIVE Method's Accuracy.** We train SigLip v2 ViT-B-16 with four triplet datasets, Open Images, WebLI, CC3M, and ours. Ours achieves broader improvements across benchmarks. While OI showed no gain, WebLI increases OCR, and CC3M offered slight improvements on some tasks, our approach highlights the benefit of using LLMs to overcome traditional data limitations for training transferable vision encoders.

such as recognizing that an ink color is not visible, rather than defaulting to the image's dominant red color (baseline vision-only embedding). Such fine-grained understanding and contextual inference were previously unattainable with vision-only embeddings.

4.3 Analysis

Impact of Training Data. We benchmarked our model by training it individually on established vision-language datasets—Open Images (Piergiovanni et al., 2022), WebLI (Wang et al., 2025), and CC3M-VQA (Changpinyo et al., 2022)—and on our novel Imagenet triplet dataset. For pre-existing datasets lacking explicit textual queries (unlike CC3M-VQA, which uses rule-extracted query-answer pairs), we employed generic queries (e.g., ”caption the image”) for a comparable training setup. As shown in Figure 7, models trained with our Imagenet triplet dataset significantly outperform those trained on existing image-language datasets across diverse benchmarks. These results strongly suggest that the scarcity of large-scale, diversified, and targeted image-query-answer data has been a key bottleneck for advancing instructed vision embeddings. While prior work often fix vision model and improves LLM, we reverse this paradigm and show the power of LLM can also facilitating more effective training of vision models.

Impact of Vision Encoder Size. As detailed in Table 3, we varied the vision encoder from ViT-T (5.4M parameters) to ViT-B (86.6M) and SO400M (891M). While larger model sizes generally yield better performance, even the compact ViT-T model achieves reasonable accuracy, which shows potential to be deployed on edge devices.

Attention Map Visualization. Figure 6 visualizes how language instructions modulate visual attention. We use heatmap to plot the attention from our language input token to the visual tokens. We contrast a baseline vision-only transformer (SigLip ViT-SO-14, left) with ours (right). Given the same input image (e.g., an apple with an ”iPod” label), the baseline's attention is instruction-agnostic since it does not take instructions. In contrast, LIVE dynamically adjusts its focus: when instructed to find ”the text”, attention localizes on the ”iPod” label; when asked for ”the fruit”, attention isolates the apple itself. This demonstrates that LIVE learns to steer its visual processing based on the language query, enabling focused computation on instruction-relevant regions.

Generalization to Out-of-Distribution (OOD) Instructions Groups. To perform a stricter test of generalization, we train and evaluate our model on semantically disjoint instruction groups. This introduces a more significant distributional shift than the deduplication used in prior experiments. The results in Table 4 show robust performance even in this challenging OOD setting.

5 Conclusion

We introduce a new paradigm for vision representation: instructing the encoder with knowledge from language models. Reversing the typical workflow of freezing a generic vision encoder, we demonstrate that injecting task-specific guidance directly into the visual system yields significant benefits. Our approach produces an efficient, lightweight encoder that enhances perceptual precision

Training Groups	SVD	FVD	FSD	FSV	FSVD
Testing Groups	F	S	V	D	FSVD
SigLIP 2 ViT-B/16	74.05	82.48	83.40	83.28	86.93

Table 4: **Leave-One-Group-Out Generalization.** To test generalization to novel instruction types, we partition our data into four categories: Fundamental Properties (F), Spatial-Textual (S), Viewpoint (V), and Dynamic Reasoning (D). We train the model while holding out each category in turn, demonstrating LIVE's ability to generalize to semantically distinct, unseen instructions.

and mitigates hallucinations without costly retraining. Our findings suggest that the key to advancing vision models on targeted tasks lies not just in scaling them, but in making them instruction-aware.

6 Ethics Statement

Our work introduces instruction-aware vision encoders that accept natural-language task specifications. While this can reduce hallucination and improve task precision, it also raises ethical considerations: Instruction following could be repurposed for harmful objectives (e.g., surveillance, targeted profiling). In our training, we do not include any harmful objectives, therefore the risk shall be minimized in our model perspective.

7 Reproducibility Statement

Code and data will be released upon acceptance, and the paper already contains the implementation details needed for reproduction. If any discrepancies arise, we will update the repository with clarifications and minimal patches.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*, 2022.

Haozhe Chen, Junfeng Yang, Carl Vondrick, and Chengzhi Mao. Invite: Interpret and control vision-language models with text explanations. In *The Twelfth International Conference on Learning Representations*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024b.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

Ainaz Eftekhar, Kuo-Hao Zeng, Jiafei Duan, Ali Farhadi, Ani Kembhavi, and Ranjay Krishna. Selective visual representations improve convergence and generalization for embodied ai. *arXiv preprint arXiv:2311.04193*, 2023.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Gemini. Gemini 2.0 flash models — vertex ai, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024.

Ouguzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pp. 113–132. Springer, 2024.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Weicheng Kuo, AJ Piergiovanni, Dahun Kim, Xiyang Luo, Ben Caine, Wei Li, Abhijit Ogale, Luowei Zhou, Andrew Dai, Zhifeng Chen, et al. Mammut: A simple vision-encoder text-decoder architecture for multimodal tasks. *Transactions on Machine Learning Research*, 2023.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7005–7015, 2021.

Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14846–14855, 2023.

Xudong Lin, Manling Li, Richard Zemel, Heng Ji, and Shih-Fu Chang. Training-free deep concept injection enables language models for video question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22399–22416, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4015–4025, 2025.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

Chengzhi Mao, Revant Teotia, Amrutha Sundar, Sachit Menon, Junfeng Yang, Xin Wang, and Carl Vondrick. Doubly right object recognition: A why prompt for visual rationales. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2722–2732, 2023.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

Sachit Menon, Ishaan Preetam Chandratreya, and Carl Vondrick. Task bias in vision-language models. *arXiv preprint arXiv:2212.04412*, 2022.

Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Pre-training image-language transformers for open-vocabulary tasks. *arXiv preprint arXiv:2209.04372*, 2022.

Michael I Posner. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25, 1980.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Sirnam Swetha, Jinyu Yang, Tal Neiman, Mamshad Nayeem Rizve, Son Tran, Benjamin Yao, Trishul Chilimbi, and Mubarak Shah. X-former: Unifying contrastive and reconstruction learning for mllms. *arXiv preprint arXiv:2407.13851*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim M Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners. *Advances in Neural Information Processing Systems*, 37:116355–116387, 2024.

Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pretraining to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*, 2025.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24884–24894, 2025.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022a.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*, 2024.

A Appendix

A.1 Limitations

Our approach enhances the controllability of visual representations using language instructions. However, its practical application and further development are subject to certain limitations, which also open avenues for future research.

Optimizing Query Design for Downstream Tasks: A primary challenge lies in the formulation of effective textual queries to maximize performance on specific downstream applications. The process of identifying the optimal phrasing, level of detail, and linguistic structure for queries that elicit the desired visual representation changes remains an empirical endeavor. It may require significant tuning for each new task or dataset. This is compounded by the inherent ambiguity and richness of natural language, where subtle variations in a query can lead to different steering outcomes, not all of which may be beneficial for the target application's accuracy. We conducted initial experiments in Figure 12, yet a principled way to discovery effective prompt is still missing.

Handling of Complex and Compositional Queries. The reliance on a pretrained text encoder constrains the complexity of queries our method can effectively interpret. Current pretrained text encoders, while powerful, often struggle with deeply compositional or abstract textual prompts. Their encoding of nuanced relationships between multiple concepts, or negation, might not be robust. Our method, therefore, performs best with relatively simple, direct queries.

Potential for Undesired Steering Outcomes. Depending on how the users provide the instructions, the model has a risk of generating biased, unsafe, or undesirable content.

A.2 Future Work

Principled Query Optimization and Discovery: Developing systematic methods or even learnable components to automatically discover or refine queries for optimal downstream performance would significantly enhance usability. This could involve techniques from prompt engineering, reinforcement learning, or semantic search to bridge the gap between user intent and effective query formulation.

Enhancing Complex Query Understanding: Future work should focus on strategies to decompose complex textual queries into simpler, manageable sub-queries that our current framework can process. Alternatively, exploring new architectures or fine-tuning regimes for the text encoder to better handle compositional semantics and logical operations directly within the query embedding space would be a valuable pursuit. This could involve incorporating structured knowledge or symbolic reasoning alongside neural representations.

Visual Grounding with Instructions: Our method can mitigate visual hallcuinations, which can be used as a component in RAG systems, to help verify, ground the reasoning and prediction of LLM.

Language-Instructed Vision Generation: Our method is a language-instructed vision encoder, which can be used as the backbone that encode semantic information in generative models, such as Diffusion. For example, by using language-instructed vision embeddings, one can train image editing models based on the instructions.

A.3 Broader Impact

Our research on language-steered vision embeddings has the potential for considerable positive societal impact, primarily by offering novel approaches to creating more equitable and robust AI systems. By enabling vision embeddings to be guided by language instructions, we introduce a mechanism for actively mitigating biases present in training datasets. This zero-shot bias mitigation capability is a significant step towards fairer AI representations, as it allows for targeted adjustments without the need for extensive retraining or dataset curation, making the development of equitable models more accessible.

Furthermore, our method enhances the robustness of vision embeddings. By training models on detailed, instructed triplets, they learn to capture nuanced, fine-grained signals from an image, moving beyond a single, holistic embedding. This improved granularity can lead to models that are more adaptable and less susceptible to being misled by irrelevant or superficial features. An important application of this enhanced instructional control is the ability to direct the model to defend against typographical attacks. This contributes to making vision models safer and more resilient to adversarial manipulations aimed at ”jailbreaking” or deceiving them.

However, we also recognize potential negative societal impacts. The same linguistic steerability that allows for bias mitigation and robustness enhancement could, if misused, be employed to intentionally introduce or amplify biases. A malicious actor could craft instructions to make the vision embeddings unfairly prejudiced against certain groups or characteristics. Currently, our work does not include a mechanism to automatically discriminate between benign and malicious instructions, nor a system to refuse potentially harmful guidance. This creates a risk of misuse, where the technology could be exploited to generate unfair or harmful representations, potentially leading to discriminatory outcomes if deployed in sensitive applications.

Future work should prioritize the development of safeguards against such misuse. This could involve research into methods for detecting and rejecting biased or malicious instructions, establishing protocols for the responsible deployment of steerable vision models, and fostering a deeper understanding of the societal implications as this technology matures.

A.4 Safeguards

Since our training set is repurposed from Imagenet dataset and other established benchmarks that has been extensively used by the field, they shall not contain image data with NSFW. For language instructions, one can implement a classifier for the instructions to classify if it is benign or malicious as a straightforward safeguard.

A.5 Pseudo Code

We provide pseudo code for implementing our LIVE encoder and training loss.

```
# Assuming text_query, image, text_answer are input batches
# Assuming t (temperature) and b (bias) are parameters
# Models: text_query_model, image_model, text_answer_model

# Precomputed query:
_zquery_raw, out_query = text_query_model(text_query)
zquery = jax.lax.stop_gradient(_zquery_raw)

# Image embeddings steered by query:
zimg, out_img = image_model(image, query_tokens=zquery)

# Text answer embeddings:
_ztxt_raw, out_txt = text_answer_model(text_answer) # **kw omitted
ztxt = jax.lax.stop_gradient(_ztxt_raw)

# Compute logits:
logits = jnp.dot(zimg, ztxt.T) # (batch_size, batch_size)
logits = logits * t + b

# Contrastive loss calculation:
batch_size = zimg.shape[0]
eye = jnp.eye(batch_size)
m1_diag1 = -jnp.ones_like(logits) + (2 * eye)

loglik = jax.nn.log_sigmoid(m1_diag1 * logits)
nll = -jnp.sum(loglik, axis=-1) # NLL per sample
loss = jnp.mean(nll) # Average loss for the batch
```

Figure 8: Pseudo JAX code for language-steered vision embedding model.

```
# ViT Input: Image + Language Query Tokens (Concise)
# Assumes: self (Flax Module), nn (flax.linen), jnp (jax.numpy)
# Config: self.T, self.dtype_mm, self.width, self.patch_size, self.posemb
# Helper: get_posemb() for positional embeddings

# 1. Image to Patch Embeddings
img_in = jnp.asarray(image, dtype=self.dtype_mm)
patches = nn.Conv(features=self.width,
                  kernel_size=(self.patch_size, self.patch_size),
                  strides=(self.patch_size, self.patch_size),
                  padding="VALID", name="patch_conv",
                   ↪  dtype=self.dtype_mm)(img_in)
n, h, w, c = patches.shape
patch_emb = jnp.reshape(patches, (n, h * w, c))
# Add positional embeddings to patch embeddings
patch_emb += get_posemb(self, self.posemb, (h,w), c, "patch_pos",
 ↪  patch_emb.dtype)

# 2. Process Query Tokens
# query_tokens input, e.g., (batch, query_feat_dim)
q_proj = nn.Dense(features=c * self.T, name="query_proj",
                  dtype=self.dtype_mm)(query_tokens)
q_proj = jnp.reshape(q_proj, (n, self.T, c))
q_pos_emb = self.param("query_pos_emb", nn.initializers.zeros,
                       (1, self.T, c), self.dtype_mm)
query_emb = q_proj + q_pos_emb

# 3. Concatenate query and patch embeddings for ViT Encoder
# Typically, sequence_axis=1 for (batch, seq_len, features)
encoder_input = jnp.concatenate([query_emb, patch_emb], axis=1)

# 'encoder_input' is then fed into the main ViT Encoder layers
```

Figure 9: Concise pseudo JAX code for ViT input processing with language queries. The $self.T$ is number of language tokens feed into the Vit.

A.6 Comparison with Existing Work

We list a comparison with existing vision language models in the followings, and visualize their architecture in Figure 10.

- A) CLIP Radford et al. (2021), SigLip Zhai et al. (2023), LiT Zhai et al. (2022b)
- B) Llava Liu et al. (2023), Gemma Team et al. (2025), Paligemma Beyer et al. (2024), Llama Grattafiori et al. (2024)
- C) CoCA Yu et al. (2022), Cappa Tschannen et al. (2023)
- D) VILT Kim et al. (2021)
- E) Falmingo Alayrac et al. (2022), BLIP Li et al. (2022), X-former Swetha et al. (2024)
- (F) Ours

Our work introduces the first vision-centric encoder that uses language to modulate visual computation for encoding target tasks. We address the scarcity of high-quality image, query, and answer triplet data by transferring the knowldge from LLM such as Gemini, and we demonstrate how language can directly control the vision encoder.

A.7 The Impact of Language Instructions for LIVE

Since our method allows feeding text instructions to the vision encoder, we have the potential to serve the final task better by improving the query. We investigated the impact of prompt text on the ImageNet classification accuracy of our SigLIP So400M model variant. We show the classification

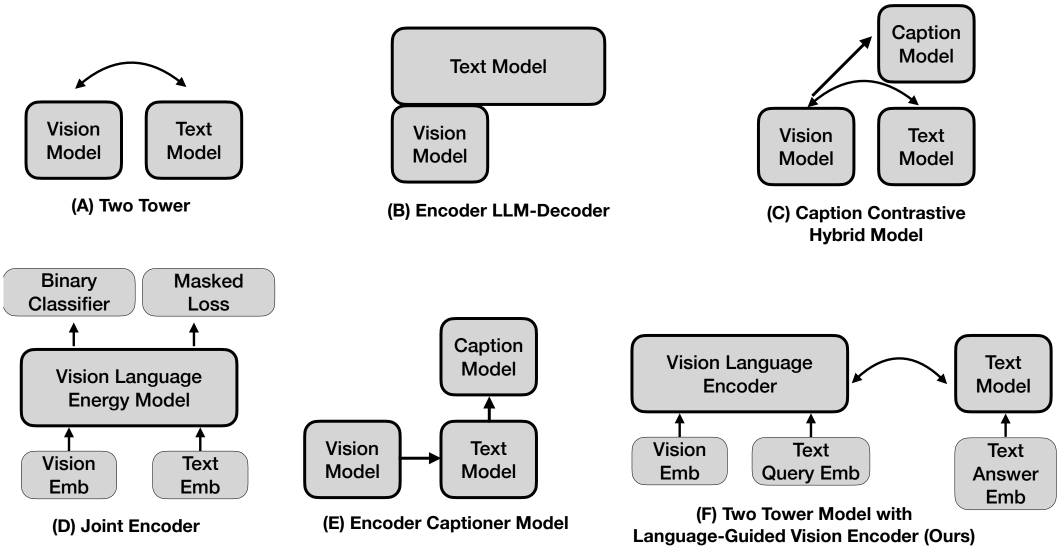


Figure 10: Comparison with existing methods. Note that B, C, E requires large language model based decoders. D does not have a embedding to perform zero-shot retrieval.

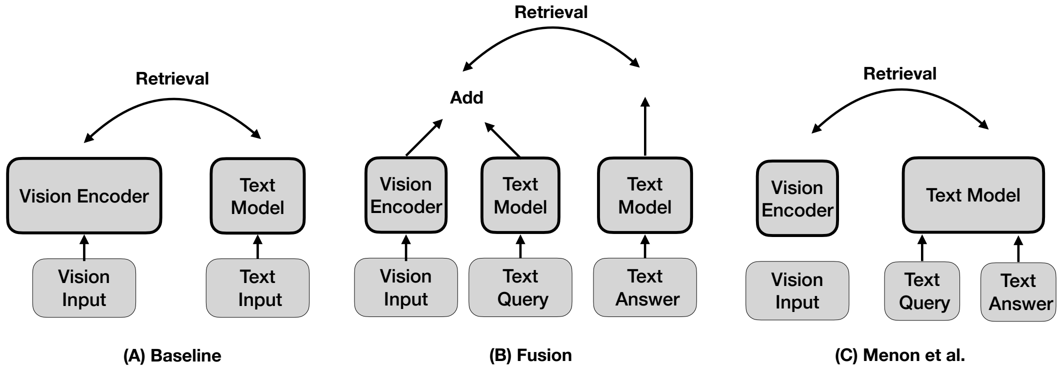


Figure 11: Illustration for baselines compared with in our paper. We take the two tower architecture (A), add the text query embedding to the embedding (B), and adding query to the text answer as description following Menon et al. (C).

accuracy of different prompt in Figure 12. Due to our model's training on more sophisticated image queries, the ImageNet classification accuracy dropped to 49.32% when no query prompt was used in retrieval tasks. Interestingly, by leveraging Gemini to evolve and generate different text prompts Yang et al. (2023), we improved the ImageNet accuracy to 68.18% using the instruction query: ”Classify the main object.” We believe this demonstrates the potential of our instructive vision foundation model for future work in prompt optimization to achieve even higher accuracy.

A.8 Additional Experimental Results.

Results on Steering Visual Representations for Text Recognition. We repurposed the ImageNet dataset for a text recognition task by rendering text from one ImageNet category onto an image of another. A visual representation that ignores this text and instead predicts the original image's category would result in 0% accuracy. Therefore, higher accuracy directly indicates the model's ability to follow instructions and perform OCR retrieval. As shown in Table 5, our approach demonstrates significant effectiveness.

Robustness Against Typographical Attacks. Vision-language models like CLIP and SigLip are known to be vulnerable to typographical attacks, where target text is appended to an image to mislead

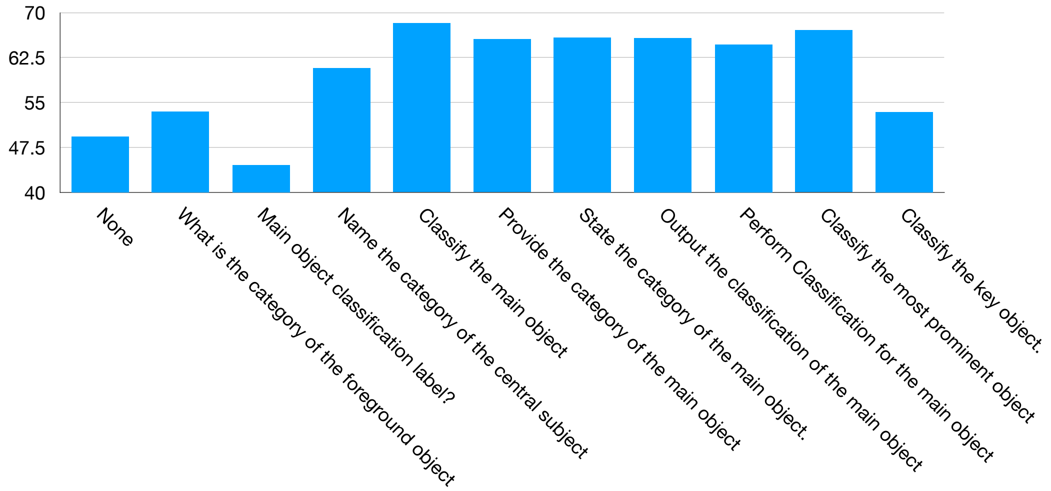


Figure 12: **Impact of Different Language Instructions for ImageNet classification task.** The y-axis shows the ImageNet classification accuracy in %. The x-axis shows the language instructions for the vision encoder. By improving the query prompts, we can improve the downstream task accuracy by up to 20 points.

	OCR Accuracy	
ViT Model	Baseline	Ours
SigLIP 2 ViT-SO-14	10.48	**38.99**

Table 5: Zero-shot accuracy to recognizing the text on Imagenet dataset. We evaluate OCR performance when text in words, potentially of different categories to the ImageNet image, is rendered in the image. If the model only perceive the original imagenet image without attending to the added text, the accuracy will be 0. While vision-only representations has a low accuracy on recognizing the text, our instructive visual embeddings allow embedding either image or text information based on instructions.

the model's representations. This vulnerability poses a significant concern for critical applications such as autonomous driving and facial authentication.

To evaluate this, we rendered a text sticker in the middle of ImageNet images, with the text explicitly stating a different class name than the original image. If the model attends to this text sticker, its accuracy drops to 0%. As shown in Table 6, baseline models exhibited a reduced ImageNet accuracy of 48.31% under these attacks. However, simply by adding the prompt, ”Ignoring text, what is the object?”, we observed a significant increase in robust accuracy, demonstrating our approach's ability to disregard typographical attacks.

	Robustness Against Typographical Attacks	
ViT Model	Baseline	Ours
SigLIP 2 ViT-SO-14	48.31	**51.48**

Table 6: **Zero-Shot test accuracy on ImageNet with typographical attacks.** When providing text sticker on top of the image, original image classification model has the tendency to be mislead by the text. By using text prompt to let the model to ignore the text, we can increase the robustness against typographical attack.

Instructive Visual Benchmark on All Language Instructions. In the main paper, we present the results on testing on unseen instructions, where we exclude all the language instructions that appear in the training. In Table 7, we also show the results on all instructions, which includes also language instructions that appear in the training. Our method consistently improves accuracy on the instructive

	ImageNet				Caltech 101			
ViT Model	SigLip	Fusion	Menon et al.	Ours	SigLip	Fusion	Menon et al.	Ours
SigLip T/14	7.84	10.85	7.93	**71.72**	7.52	6.64	9.49	**26.50**
SigLip B/16	8.45	9.35	8.17	**83.51**	8.83	7.56	12.3	**38.74**
SigLip 2 B/16	9.29	10.91	9.50	**84.54**	8.18	8.05	14.8	**37.12**
SigLip 2 So400m	9.21	10.46	9.43	**85.00**	8.08	8.18	15.04	**37.64**

	SUN				RefCOCO			
ViT Model	SigLip	Fusion	Menon et al.	Ours	SigLip	Fusion	Menon et al.	Ours
SigLip T/14	6.13	5.50	6.72	**26.87**	5.72	5.72	5.72	**33.52**
SigLip B/16	8.41	8.06	9.68	**41.55**	6.93	6.94	7.09	**47.24**
SigLip 2 B/16	10.08	10.02	14.41	**41.41**	7.75	7.75	9.72	**45.42**
SigLip 2 So400m	10.81	10.54	15.26	**44.68**	6.63	6.68	7.65	**47.80**

Table 7: **Zero-Shot Accuracy on Instructive Visual Benchmark repurposed from ImageNet, Caltech 101, SUN, and RefCOCO.** We directly test our model on these datasets without any training on them. This is in contrast to prior work that require finetuning on those downstream tasks Mao et al. (2023); Beyer et al. (2024) to do them.

visual benchmark. Despite some instructions being encountered during training, the task's difficulty persists. This is attributed to the new image and data domains, and the fact that many tasks remain non-trivial even with instruction familiarity.

Ablation Study on Cross-Instruction Generalization We investigate the ability of our learned embeddings to generalize to unseen instruction families after training on a distinct set. Utilizing Gemini, we automatically categorize ImageNet instructions into four broad families: fundamental properties (F), spatial-textual symbolic tasks (S), viewpoint composition aesthetics tasks (V), and dynamic inferential interpretive reasoning tasks (D).

Table 4 presents our results where a SigLip 2 B/16 model is trained on three of these instruction groups and evaluated on the deliberately held-out fourth group on ImageNet. While training and testing on all groups yields 86.93% accuracy, testing on our hold-out subgroups results in only a 1-2 percentage point accuracy drop for three of our studies. Notably, when not training on F (fundamental properties), the model experiences a significant accuracy drop, underscoring the importance of training on instructions related to fundamental properties.

A.9 Training Dataset

We conducted an in-depth analysis to understand the distribution of language instructions generated by our LLM for the ImageNet dataset. Our process involved two key steps: First, we used Gemini Flash 2.0 to define 66 distinct subcategories for vision-related questions, which are depicted in Figure 13. Second, we employed Gemini Flash 2.0 to assign each question within our expansive 16-million synthetic image-query-answer triplet dataset to one of these 66 categories, or to an ”others” category if it didn't fit.

The resulting distribution, visualized in Figure 13, reveals significant variations in instruction frequency. ”Material identification via Visual Properties” was by far the most common, accounting for roughly 2.2 million data entries. In contrast, ”Fractal Properties/Self-Similarity Analysis” was rarely observed, with only 140 associated queries.

A.10 Testing Dataset

A.10.1 Established Benchmarks

MMVP. In this paper, we use the MMVP-VLM benchmark, which are divided into 9 visual patterns. The benchmark consists of image pairs with corresponding answer pairs to retrieve. The original MMVP only comes with text answers, no text queires. Yet since they are divided into 9 categories with answers that has a good description for the task to ask about. We create text quries, which itself,

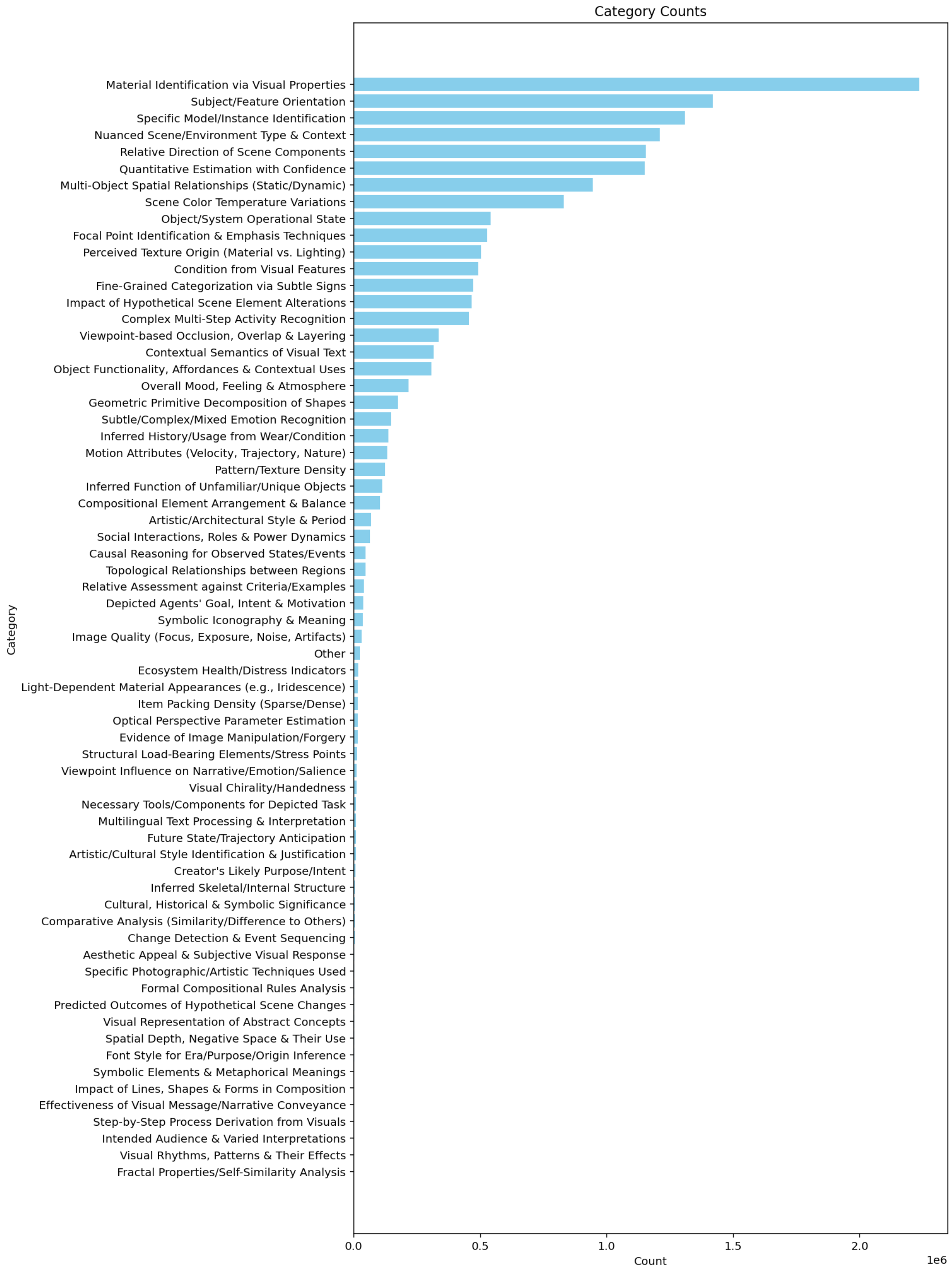


Figure 13: **The Histogram of Query Categories Generated in our Language-Instructed ImageNet.** We first use LLM to generate a taskonomy of visual queries. We then use LLM to label each instructions we generate to one of the categories. We show the counting plot. The data generated show a long tail distribution.

does not offer any additional information to distinguish the text answer, since for example, the answer pair to discriminate is ”A minion smiling with tounge out” and ”A minion smiling without tounge out”, our added query: ”Is the minion smiling with tounge out” does not offer additional information on the tounge's status, but a repeat of the answer context. Note the chance prediction accruracy on MMVP is 50% due to the binary choice.

VQA v2, GQA. While our target is to evaluate how language can steer the visual representations, as benchmarked by our above datasets that designed to evaluate this, like having rich query and answer pair for the same image. We also use existing visual question answering tasks like VQAv2 and GQA, which often has single query and ansewr for the same image. We subsample the first 500 samples from GQA to validate our approach for all our experiments. By adding our instructions, we achieve significantly higher accuracy than vanilla models. Note that for VQA task, the query tend to already contain a lot of information about the image, therefore, the Menon et al achieve higher accuracy largely due to the query allows better retrieval for the image.

A.10.2 Our Benchmarks

For the following repurposed dataset to evaluate this language-instructed vision embeddings, we use the same prompt to generate the test answer and query:

```
Provide a numbered list of interesting visual questions about the image,
    followed by the corresponding answers.
```

Note since those are unseen images, and new domains for four of them, the Gemini generate questions are often very different, allowing us to perform zero-shot evaluation on both: 1) novel data category and domain but instructions could be seen before, 2) novel data category, domain, and unseen instructions. We report the (2)'s results in the main paper due to the space limitation. We will also report the accuracy for (1) in later section.

ImageNet. ImageNet validation was originally designed for evaluating classification tasks. We repurpose it to also benchmark instructive visual embeddings. The queries are generated by gemini condition on the Image. In the main paper, we remove all instructions are appear in the training. Therefore, the numbers shown is on unforeseen, new instructions. In addition, in the appendix, we also show the retrieval on all the instructions generated without removing the ones that overlap with the training queries. There are 145549 data for the validation data in the paper after removing the ones with instructions appear in the training. Before removing the data with seeing instructions, is 551514. We retrieve the answer from 1000 answers, which contains the groundtruth and 999 random others.

Caltech101 We repurpose the test set via Gemini, to generate open quries and corresponding answers. In the main paper we remove queries that overlap with training. We also show the results for the set without removing the overlapping ones.

SUN We repurpose the test set via Gemini, to generate open quries and corresponding answers. In the main paper we remove queries that overlap with training. We also show the results for the set without removing the overlapping ones.

RefCOCO We use the images with rendered bounding box to to create the test datasets. We feed the image with bounding box to LLM to generate open quries and corresponding answers. Note that the task is zero-shot because bounding box is not given in ImageNet.

ImageNet OCR Test We render text on the ImageNet validation images, where the text are the name of a different category. Therefore, the model will have different predictions by looking at the text or the image object category itself.

A.11 Attention Visualizations

We provide more attention visualizations of our encoder in Figure 14. Guided by language instructions, without supervision on where the model shall look at, our LIVE encoder learns to focus on the part of the image that is corresponding to the language instructions.

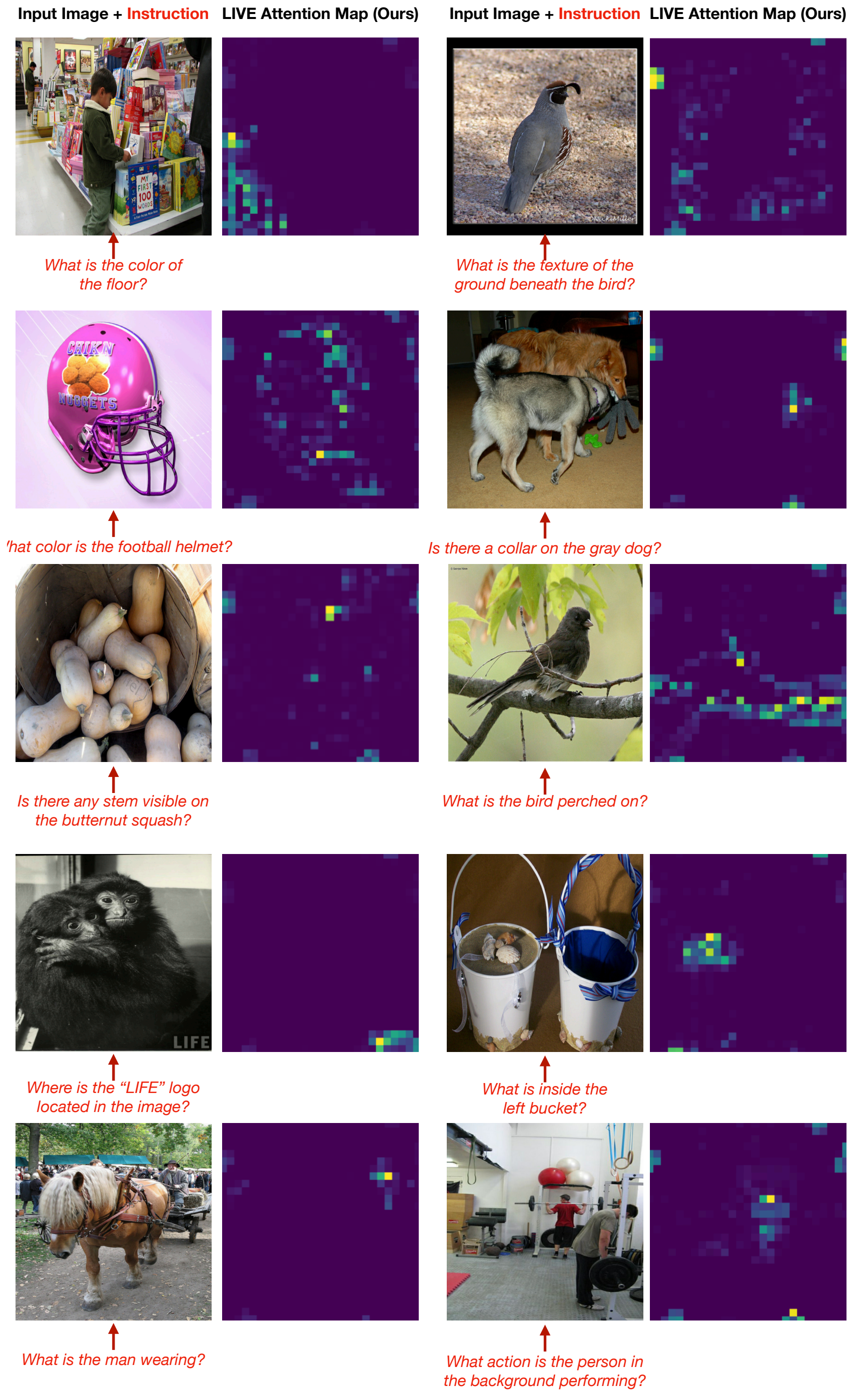

Input Image + Instruction
LIVE Attention Map (Ours)
Input Image + Instruction
LIVE Attention Map (Ours)
What is the color of the floor?
What is the texture of the ground beneath the bird?
'hat color is the football helmet?
Is there a collar on the gray dog?
Is there any stem visible on the butternut squash?
What is the bird perched on?
Where is the "LIFE" logo located in the image?
What is inside the left bucket?
What is the man wearing?
What action is the person in the background performing?

Figure 14: **Attention Visualizations of Our LIVE Encoder.** Guided by language instructions, the ViT model learn to focus on relevant parts, effectively prioritizing information and ignoring distractions. This is achieved without any direct supervision on the region the model shall focus on, showing the active, selective capabilities can be automatically learned by our encoder. Examples are randomly draw from ImageNet validation set that was not trained on.