CONCEPTS OR SKILLS? RETHINKING INSTRUCTION SELECTION FOR MULTI-MODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing instruction selection methods in vision-language learning rely on sample embeddings to guide data choice. These embeddings are typically derived from pure vision encoders or small multimodal models and they primarily capture *visual concepts* while under-representing *visual skills* such as counting, spatial reasoning, or commonsense inference. This imbalance overlooks a key distinction: multimodal benchmarks vary widely in whether they emphasize conceptual grounding or skill-based reasoning. We show that this concept–skill axis provides a systematic lens for characterizing benchmark demands, and that prioritizing one dimension often comes at the expense of the other. To address this, we introduce a simple benchmark-aware data selection framework that adapts training data to the dominant alignment factor of each benchmark. Across twelve diverse benchmarks, our approach yields consistent improvements, especially in low-data regimes (+0.9% over the best existing baseline on average and +1.2% on the skill-focused subset). More broadly, our findings highlight that advancing multimodal learning requires explicit recognition of the dual role of concepts and skills in shaping benchmark behavior.

1 INTRODUCTION

Recent progress in vision-language modeling has been driven by the modular combination of large pretrained language models (LLMs) with powerful visual encoders, typically connected via a modality adapter that transforms visual features into a format compatible with the LLM's input space (Zhu et al., 2023; Liu et al., 2023; Wang et al., 2024). This architecture allows models to leverage the linguistic capabilities of LLMs while incorporating rich visual understanding from pretrained vision backbones. Extensive continual pretraining on paired vision-language data is often required to bridge the modality gap (Li et al., 2023; Awadalla et al., 2023; Liu et al., 2024). In addition, vision-language instruction tuning has become essential for aligning the joint model's behavior with desired visual-linguistic tasks (Xu et al., 2023; Li et al., 2024; Lin et al., 2023).

Vision-language instruction tuning has emerged as a powerful framework for aligning multimodal models with human-desired behaviors, especially in tasks requiring both visual perception and linguistic reasoning. This tuning process serves two major purposes: first, it helps models learn to associate visual representations with the corresponding textual concepts (Safaei et al., 2025), and second, it enables the acquisition of new visual capabilities such as counting objects, reasoning about spatial relations, and inferring physical properties (Chen et al., 2025; Shao et al., 2024). These dual roles are critical for deploying vision-language models in real-world applications where generalization to both seen and unseen instructions is required.

Despite instruction tuning becoming a standard approach, selecting which instructions to use remains a critical challenge. Existing data selection methods typically rely on embeddings of training examples to guide the choice (Wu et al., 2024; Lee et al., 2024; Safaei et al., 2025). These embeddings, whether derived from vision encoders or small multimodal models, primarily capture visual concepts while under-representing skill-based reasoning such as counting, relational understanding, or commonsense inference. As a result, instruction selection may align well with concept-focused benchmarks but fail to prioritize the skills actually required by a task. This raises a fundamental question: Should instruction selection focus on matching the visual concepts present

in a benchmark, or the reasoning skills it requires? More concretely, can benchmark performance be improved by tuning on data that emphasizes skills, rather than just concepts?

We conduct a systematic analysis across a diverse range of vision-language benchmarks to investigate how different types of instruction similarity (concept versus skill) affect downstream performance. Our study reveals a clear pattern: some benchmarks benefit more from instructions that emphasize reasoning skills (e.g., object counting or relational tasks), while others benefit more from concept-aligned instructions (e.g., object or scene categories). This also suggests why existing embedding-based selection methods tend to prioritize concepts: the evaluation benchmarks they report are predominantly concept-focused, which implicitly biases instruction selection toward conceptual similarity.

Motivated by these findings, we propose a simple yet effective targeted instruction selection method that adapts to the nature of each benchmark. The goal of the method is not just another data selection strategy. Rather it highlights the importance of understanding the nature of benchmarks and the tradeoffs when performing data selection. Specifically, we first extract the dominant concepts and skills present in a given evaluation set using automated instruction parsing and skill/concept taxonomy alignment. We then determine whether the benchmark is concept-dominant or skill-dominant using validation performance differentials, and finally, we select training instructions that most closely align with the benchmark's dominant type. This targeted strategy allows the model to focus on the most relevant conceptual or skill-based inductive biases.

We evaluate our approach across twelve standard vision-language benchmarks, covering a wide range of task types and difficulty levels. Our results demonstrate consistent performance gains, with an average improvement of +0.9% over the strongest baseline using untargeted instruction tuning and +1.2% on the skill-focused subset. These findings underscore the importance of benchmark-aware instruction selection and open new directions for task-adaptive multimodal learning.

2 Related Work

2.1 Vision Language Model

Large language models (Dubey et al., 2024; OpenAI; Team et al., 2024; Chiang et al., 2023) have demonstrated remarkable performance across a wide variety of tasks. This success is largely attributed to pretraining on trillions of tokens, followed by post-training techniques such as reinforcement learning with human feedback (Christiano et al., 2017). Building on their capabilities in the text domain, recent research has extended LLMs to handle additional modalities such as images. MiniGPT-4 (Zhu et al., 2023) integrates a pretrained Vision Transformer (ViT) backbone with a Q-Former and a single linear projection layer, combining it with the Vicuna language model to achieve strong performance on multimodal tasks. Concurrently, InstructBLIP (Dai et al., 2023) employs a similar approach to generate instruction-following responses conditioned on both images and text prompts. LLaVA (Liu et al., 2023; 2024) converts a large language model into a multimodal model by first encoding images with a CLIP (Radford et al., 2021) encoder, then mapping the visual features into the text embedding space through a linear MLP, forming a simple yet effective integration strategy.

2.2 Visual Instruction Data Selection

Dataset selection (Har-Peled & Mazumdar, 2004; Roux et al., 2012; Campbell & Broderick, 2018; Mirzasoleiman et al., 2020) has been extensively explored to enhance the training efficiency of models. Recently, this line of research has been extended to vision-language models, aiming to reduce training costs while preserving performance. Coincide (Lee et al., 2024) partitions the dataset into numerous subsets and retains samples based on the transferability of clusters. ICONS (Wu et al., 2024) identifies important samples by measuring the influence of individual data points, defined via gradient similarity with a validation set. Prism (Bi et al., 2025) introduces a training-free method that utilizes Pearson correlation analysis to measure the intrinsic visual encoding capabilities of MLLMs. In contrast, PreSel (Safaei et al., 2025) approaches the problem differently: it first applies a filtering mechanism to identify high-quality images, and only then generates instructions for those selected samples.

Despite significant progress in achieving strong performance with reduced data, there is limited research on how different tasks are affected by underlying concepts or skills. We observe that the benchmarks they report performance on generally require more concept understanding capabilities (e.g. simple yes/no questions asking for the existence of objects), leading their methods to be more biased towards concept-focused examples. When evaluating on skill-heavy benchmarks, we notice these data selection methods experience a significant drop in their performance edge compared to the random baseline, further validating the concept bias in the selection strategies (see Table 1).

2.3 Concepts vs Skills

We distinguish sharply between the dichotomy: *concepts*, which refer to the visual entities, attributes, and objects present in an image, i.e., **what** appears, and *skills*, which encapsulate the reasoning operations or judgment strategies necessary for correctly interpreting or answering questions about those entities, i.e., **how** to analyze them). This distinction echoes the formal view of compositionality in VQA proposed by Whitehead et al. (2021), who explicitly model a skill–concept decomposition to enable generalization across unseen combinations of skills (e.g. "color" judgment) and concepts (e.g. "car"). While concepts ground models in visual content, skills capture the latent reasoning patterns—such as counting, spatial inference, or trend analysis—that drive downstream task success. Coincide (Lee et al., 2024) integrates concept and skill to jointly define a notion of similarity, which is then used to maximize diversity in the selected samples. In our case, decoupling these two axes allows selection methods to target training examples based on what a model needs to recognize versus what it needs to reason about, facilitating more precise alignment with the cognitive demands of various vision–language benchmarks.

3 Methodology

We introduce a retrieval-based framework to compare *concept-prioritized* and *skill-prioritized* selection strategies for curating vision-language instruction-tuning data. The approach explicitly separates the notion of *visual concept* from that of *visual skill*. For a given instruction, similar examples are retrieved from a candidate pool using nearest-neighbor search performed either in the concept space or in the skill space.

3.1 Problem Formulation

Let $\mathcal{I} = \{(x_i, v_i, y_i)\}$ denote a pool of multi-modal instruction examples, where each x_i is a set of natural language instructions, v_i is one or more associated images, and y_i is the expected responses. Given a target set of downstream tasks $\mathcal{T} = \{T_1, \ldots, T_K\}$, our objective is to select a subset $\mathcal{I}^* \subset \mathcal{I}$ that maximizes performance on the downstream tasks after instruction tuning. Each task T_k corresponds to a benchmark dataset characterized by a specific input-output format and evaluation metric. We hypothesize that the most beneficial instruction subsets vary by task, and that alignment between the instructions' visual content v_i and the task's demands plays a crucial role in downstream generalization.

3.2 Data Representation

Concept Representation. For each instruction example $(x_i, y_i) \in \mathcal{I}$, let v_i denote the image paired with the instruction x_i. We obtain a visual embedding $c_i \in \mathbb{R}^d$ for v_i by passing it through a pretrained vision encoder $\phi(\cdot)$:

$$c_i = \phi(v_i).$$

The representation c_i serves as a *concept embedding*, capturing the semantic content of the visual modality while remaining agnostic to the textual instruction x_i. Concept embeddings provide a task-independent characterization of the visual domain of $\mathcal{I}$ and allow us to measure similarity between instructions based on their associated visual content. These embeddings form the basis for concept-targeted selection strategies, where subsets $\mathcal{I}^* \subset \mathcal{I}$ are chosen to align the visual coverage of selected instructions with the demands of downstream tasks $\mathcal{T}$.

Skill Representation. The main technical contribution of this work lies in the construction of a *skill representation*. Unlike concepts, visual skills are not directly annotated in $\mathcal{I}$ and must be inferred. To address this, we introduce an automated pipeline that associates each instruction example $(x_i, v_i, y_i) \in \mathcal{I}$ with a skill embedding $s_i \in \mathbb{R}^m$:

1. **Skill isolation through large language models.** For each triplet (x_i, v_i, y_i), a large language model is prompted with the query: *"What visual skills are required to answer this instruction correctly? [[x_i]]"* The response is a concise natural language description, denoted σ_i, that may mention multiple skills (e.g., "object counting and spatial reasoning").
2. **Skill embedding extraction.** The skill description σ_i is converted into a fixed-dimensional vector using a pretrained sentence embedding model $\psi(\cdot)$:

$$s_i = \psi(\sigma_i).$$

The resulting representation s_i explicitly decouples *what an image depicts* (captured by concept embeddings c_i) from *the reasoning skills required* to interpret it. This enables skill-targeted selection strategies, in which subsets $\mathcal{I}^* \subset \mathcal{I}$ are chosen based on alignment between the inferred skill requirements of instructions and the demands of downstream tasks $\mathcal{T}$.

3.3 Nearest-Neighbor Data Selection

To curate a relevant and targeted subset of data tailored to a given query instruction, we employ a nearest-neighbor retrieval strategy. The method operates within high-dimensional embedding spaces, identifying and selecting data points that exhibit the highest similarity to the query based on a chosen distance metric (e.g. cosine similarity). Our approach is distinguished by its application of this retrieval mechanism across two distinct, semantically meaningful vector spaces: a **concept space** and a **skill space**. This dual-framework allows for the nuanced selection of examples based on different criteria of relevance.

Through the independent application of these two distinct strategies for each query, we generate two unique subsets of data. Although drawn from the same master data pool, these subsets are purposefully curated based on orthogonal principles. This dual-selection framework is foundational to our methodology, enabling a controlled analysis of how the nature of retrieved data influences downstream task performance and model behavior.

3.4 Downstream Evaluation

To empirically validate our proposed data selection framework, we conduct a series of downstream evaluations designed to systematically measure the impact of our concept-driven versus skill-driven data curation strategies. Crucially, this setup constitutes a controlled experiment. We perform two parallel fine-tuning runs where all factors, including the model, learning rate, and hyperparameters are held constant. The sole differentiating variable is the dataset used for tuning: one curated via concept-prioritized selection and the other via skill-prioritized selection. This ensures that any observed variance in downstream multimodal performance can be directly and confidently attributed to the data curation strategy itself.

The efficacy of each approach is then assessed by evaluating the two resulting models on a comprehensive suite of downstream benchmarks. We adopt the following naming convention for the two models: "Concept↑" refers to the model instance that was fine-tuned on the dataset of nearest neighbors selected from the concept space. "Skill↑" refers to the model instance that was fine-tuned on the dataset of nearest neighbors selected from the skill space.

4 Experiments

We evaluate our proposed instruction selection strategies across twelve diverse vision-language benchmarks and two instruction datasets. Our goal is to measure the effectiveness of concept- and skill-targeted instruction selection in comparison with several untargeted baselines under data budget constraints.

4.1 Experimental Setup

Dataset We conduct training budget constraint experiments on two datasets: LLaVA-1.5 (Liu et al., 2024) with 665k examples and ALLaVA-4V (Chen et al., 2024) with 1.2M examples. Each example includes one or more instruction-response pairs and an associated image. We experimented with sampling budgets of 5% and 10% for the LLaVA-1.5 instruction pool and 2.5% for the ALLaVA-4V instruction pool.

Baselines We compare concept-prioritized (`Concept`↑) and skill-prioritized (`Skill`↑) targeted selection strategies against three untargeted baselines: **Random** sampling, **Coincide** (Lee et al., 2024), which selects samples maximizing diversity measured via the embeddings of a smaller MLLM, and **PreSel** (Safaei et al., 2025), which relies on the diversity defined by the image embedding and the utility with respect to downstream tasks for instruction selection. Notably, although ICONS (Wu et al., 2024) is also a relevant baseline for targeted selection, we were unable to reproduce their results due to the high computation cost associated with caching the LoRA gradients for every training instance.

Training details We implement instruction tuning using the LLaVA-1.5 framework, which integrates a CLIP-ViT image encoder with a Vicuna-7B language model. Fine-tuning is performed using LoRA adapters for parameter efficiency. All models are trained for one epoch using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 128, and an image resolution of 224×224 pixels. Training is distributed across four NVIDIA A6000 GPUs. Performance for methods with reported standard deviation is averaged over three random seeds.

Embedding details We precompute embeddings for all instruction examples and benchmark samples. We use the FAISS library for efficient nearest neighbor search to identify top-k most similar instruction samples per benchmark. Concept embeddings are obtained using a zero-shot CLIP image feature extractor on the images. Skill embeddings are obtained using first prompting GPT-4o with the questions in an instruction for the relevant visual skills required to answer the question. The skill descriptions are then encoded with an open-sourced sentence transformer, MiniLM-L6-v2, to extract a 384-dimensional dense embedding.

4.2 Evaluation Protocol

We evaluate each model in a zero-shot setting on the downstream benchmarks, reporting task-specific metrics such as accuracy (GQA, ScienceQA) and exact match (TextVQA, OCR-VQA) where applicable. The random baseline and other baselines for the 5% LLaVA setting are repeated with three random seeds to account for training variability, and we report the average performance along with standard deviations. For other settings, we report the results of one experiment run due to computation constraints. The standard deviations from the random baseline are utilized to determine if other baselines outperform random statistically significantly.

We evaluate our methods on twelve benchmarks: VQAv2, GQA, VizWiz, ScienceQA (SQA-I), TextVQA, POPE, MME, MMBench (en), LLaVA-Bench, AI2D, OK-VQA, and ST-VQA. These benchmarks encompass a variety of tasks, including general VQA, OCR, and scientific reasoning.

4.3 Experiment Results

Table 1, 2, and 3 present the comparisons of the baselines on different benchmarks. The benchmarks are split into three sections separated by horizontal lines: the top section corresponds to concept-targeted tasks, while the bottom section corresponds to skill-targeted ones. The middle section consists of hybrid benchmarks that benefit similarly from concept knd skill targeting. The best performing method is highlighted in bold, and the second is underlined.

LLaVA 5% Table 1 presents results under the 5% LLaVA budget. In this low-resource setting, instruction selection is particularly impactful. Targeted strategies yield strong gains on most benchmarks, especially the skill-focused tasks with specific domain characteristics or reasoning.

LLaVA 10% Table 2 shows results with a 10% selection budget. While increased data volume reduces the performance gap between methods, targeted selection still provides notable improvements on benchmarks with strong domain or skill alignment. Notably more benchmarks across the board benefit from skill-focused selection. We conjecture this is a consequence of learning signals from skill-neighbors being less likely to saturate.

ALLaVA 2.5% Table 3 reports results from the 2.5% ALLaVA setting, where instructions are more diverse and noisier. In this setting, the targeted methods significantly outperform the random baseline, highlighting the importance of instruction quality and alignment in low-data regimes, especially for larger and more diverse instruction datasets.

Table 1: Experiment results for 5% data selection on LLaVA-1.5.

Category	Benchmark	Untargeted			Targeted (Ours)		
		Random	Coincide	PreSel	Concept↑	Skill↑	C-S
Concept	VizWiz	28.4 ± 0.7	28.7	28.4	**30.2**	28.6	+1.6
	VQAV2	72.2 ± 0.2	**73.3 ± 0.0**	72.3 ± 0.2	72.1 ± 0.0	71.7 ± 0.1	+0.5
	TextVQA	52.0 ± 0.3	50.3 ± 3.3	51.0 ± 0.1	**54.8 ± 0.3**	54.0 ± 0.5	+0.8
	GQA	52.7 ± 0.5	53.5 ± 0.1	51.8 ± 0.5	**54.0 ± 0.2**	53.5 ± 0.5	+0.5
	MME	1259.3 ± 12.8	**1340.8 ± 2.2**	1290.9 ± 43.5	1296.4 ± 5.3	1287.7 ± 34.2	+8.7
Hybrid	POPE	84.3 ± 0.6	83.5 ± 0.3	83.8 ± 0.6	84.1 ± 0.5	**84.0 ± 0.7**	+0.1
	STVQA	44.7 ± 0.1	46.5 ± 0.1	45.5 ± 0.4	46.9 ± 0.1	**47.2 ± 0.2**	-0.3
	LlaVa-Bench	66.7 ± 1.4	**67.8 ± 0.5**	66.0 ± 2.3	66.3 ± 1.9	67.5 ± 0.2	-1.2
Skill	MMBench(en)	55.8 ± 1.4	55.1 ± 0.2	55.4 ± 1.4	56.4 ± 0.8	**57.6 ± 1.5**	-1.2
	SQA-I	65.9 ± 0.5	66.3 ± 0.2	67.0 ± 0.7	65.7 ± 0.1	**67.6 ± 1.3**	-1.9
	AI2D	50.8 ± 0.6	50.3 ± 0.1	51.0 ± 0.4	49.2 ± 0.4	**53.0 ± 2.7**	-3.8
	OK-VQA	45.2 ± 0.7	**51.0 ± 0.9**	37.8 ± 4.5	43.2 ± 1.7	48.0 ± 2.7	-4.8

Table 2: Experiment results for 10% data selection on LLaVA-1.5.

Category	Benchmark	Untargeted			Targeted (Ours)		
		Random	Coincide	PreSel	Concept↑	Skill↑	C-S
Concept	VizWiz	28.8 ± 1.1	29.7	29.8	29.2	**30.8**	-1.6
	VQAV2	74.0 ± 0.2	**75.0**	74.0	74.5	73.6	+0.9
	TextVQA	53.5 ± 0.7	53.4	53.3	55.0	**55.7**	-0.7
	GQA	56.0 ± 0.2	56.6	56.0	56.6	**57.2**	-0.6
	MME	1349.6± 34.1	1382.4	**1387.3**	1368.6	1302.7	+65.9
Hybrid	POPE	84.0 ± 1.0	84.3	84.3	**84.9**	84.3	-0.6
	STVQA	47.1 ± 0.6	48.2	47.9	47.7	**48.7**	-1.0
	LlaVa-Bench	67.0 ± 3.2	**68.6**	66.9	68.4	68.2	+0.2
Skill	MMBench(en)	58.1 ± 0.8	**60.8**	57.7	57.7	59.5	-1.8
	SQA-I	67.7 ± 0.5	66.9	66.0	67.1	**70.4**	-3.3
	AI2D	52.2 ± 0.5	53.3	51.5	51.9	**54.2**	-2.3
	OK-VQA	49.5 ± 1.3	**53.7**	50.4	50.7	51.9	-1.2

5 Discussion

5.1 Different Prioritization Benefits Different Downstream Tasks

Our results indicate that the effectiveness of data selection strategies depends strongly on the nature of the benchmark. *Concept-prioritized selection* tends to benefit benchmarks where success depends primarily on recognizing and localizing objects within an image. These tasks, such as VQAv2, VizWiz, and MME, are dominated by relatively straightforward yes/no or short-answer questions that can be answered once the relevant visual elements are correctly identified. In contrast, *skill-prioritized selection* shows clear advantages on benchmarks that demand reasoning or specialized visual competencies beyond object recognition. Datasets such as SQA-I, OK-VQA, and AI2D require abilities like commonsense reasoning, fine-grained attribute discrimination, reading embedded text (OCR), or multi-step inference over visual evidence. The divergence in performance suggests that concept-driven and skill-driven selection are complementary: the former strengthens a model's ability to ground answers in visual content, while the latter enhances its ability to execute more complex reasoning over that content.

5.2 Untargeted Baselines Implicitly Trade Off Performance

While untargeted baselines such as Coincide and PreSel perform well on average, they tend to make implicit tradeoffs across tasks. These methods favor globally frequent instruction patterns, which may benefit general-purpose benchmarks like VQAv2 or GQA but degrade performance on domain-specific or skill-intensive tasks. Several concept-centric benchmarks are not reported in the original

Table 3: Results for ALLaVA 2.5% data selection.

Category	Benchmark	Untargeted	Targeted (Ours)		
		Random	Concept↑	Skill↑	C-S
Concept	VizWiz	21.0 ± 0.4	**31.1**	30.5	+0.6
	VQAV2	52.5 ± 2.8	**72.3**	70.9	+1.4
	TextVQA	36.6 ± 3.2	52.2	**52.8**	-0.6
	GQA	34.8 ± 1.9	**52.7**	51.7	+1.0
	MME	854.4 ± 97.8	1208.5	**1222.9**	-14.4
Hybrid	POPE	76.1 ± 1.6	**83.3**	82.3	+1.0
	STVQA	29.8 ± 3.5	46.8	**47.5**	-0.7
	LlaVa-Bench	63.2 ± 2.5	**76.6**	70.7	+5.9
Skill	MMBench(en)	29.9 ± 1.8	52.0	**54.8**	-2.8
	SQA-I	44.9 ± 5.4	62.1	**66.5**	-4.4
	AI2D	42.4 ± 2.6	50.3	**54.0**	-3.7
	OK-VQA	0.4 ± 0.3	**38.4**	24.7	+13.7

baseline studies, and we observe that performance on such unreported tasks tends to suffer when these heuristics are applied. This highlights the importance of understanding the hidden biases and tradeoffs that come with untargeted selection. Another consequence of this global-averaging approach is that special cases such as TextVQA or SQA, which require more specialized skills (e.g., OCR, compositional reasoning), are easily underrepresented and neglected. Our findings suggest that incorporating a stronger focus on skill diversity into otherwise untargeted selection strategies may help mitigate these limitations, reducing the cost of optimizing for average performance while still supporting tasks that fall outside the dominant distribution.

5.3 Hybrid Selection Strategies

We also investigated whether combining concept-targeted and skill-targeted strategies could provide the best of both worlds. For each benchmark, we computed a relevance score for each instruction with respect to visual concepts and skills, and then explored multiple methods for combining these scores into a unified selection criterion. Specifically, we experimented with (1) summing the concept and skill relevance scores, (2) taking the maximum of the two scores, and (3) splitting the selection budget evenly so that half of the instructions were chosen based on concept relevance and the other half on skill relevance.

Table 4: Experiment results for 5% data selection on LLaVA-1.5 for comparing hybrid strategies for combining concept and skill signals. Notice that no hybrid methods consistently outperform the concept or skill only baselines.

Category	Benchmark			Concept-Skill Hybrids		
		Concept↑	Skill↑	Max	Sum	Split
Concept	GQA	<u>54.0</u>	53.5	53.9	**54.9**	53.7
	MME	1296.4	1287.7	1312.3	<u>1312.5</u>	**1337.2**
Skill	SQA-I	65.7	67.6	**70.1**	<u>69.3</u>	68.1
	OK-VQA	43.2	**48.0**	41.9	<u>47.4</u>	<u>47.4</u>

Table 4 presents the three hybrid approaches on a representative concept- and skill-targeted subtasks. Surprisingly, none of these hybrid strategies consistently outperformed the single-targeted approaches. In nearly all benchmarks, the worst hybrid variant underperformed the better of the two targeted strategies. This suggests that combining the two signals indiscriminately dilutes the effect of the dominant alignment factor, whether it be concept or skill, and there is no straightforward method to directly incorporate the concept and skill signals. Our findings imply that benchmarks tend to benefit strongly from one type of alignment at a time rather than a naive mixture of both.

5.4 Predicting Benchmark Alignment via Mutual Ranking

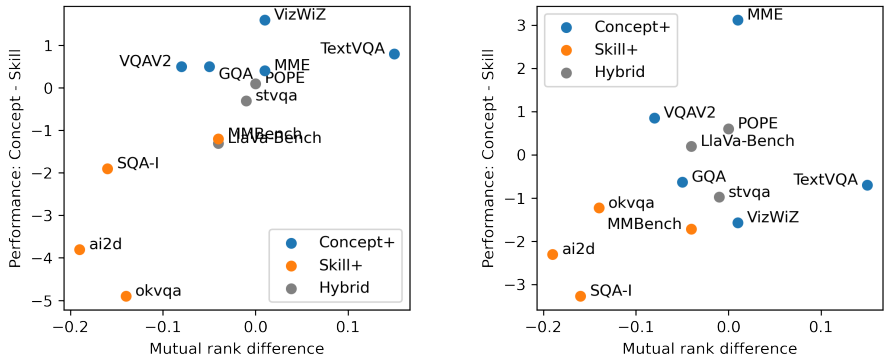
The unsuccessful attempt to achieve the best-of-both-worlds reinforces the importance of being able to predict *which* dimension—concept or skill—is more relevant for a given benchmark. Rather than attempting to combine the two types of selection heuristics, learning to automatically choose between them appears to be a more effective path forward. We explore whether it is possible to predict which alignment factor dominates a benchmark using a simple mutual ranking analysis. The goal is to infer, without running two separate fine-tuning experiments, whether a benchmark is likely to benefit more from concept-targeted or skill-targeted selection.

Method. For a given benchmark, we first construct two relevance rankings over the entire instruction pool. The *concept ranking* ranks samples by their visual concept similarity using c_i to the benchmark images, and *skill ranking* ranks by their visual skill similarity using s_i.

To capture the relationship between these two rankings, we examine the top-1 nearest neighbor in each list and measure where that sample appears in the opposite ranking. Concretely, for the top-1 skill-ranked sample, we record its rank in the concept ordering, and for the top-1 concept-ranked sample, we record its rank in the skill ordering. We refer to these two cross-ranks as $R_{c|s}$ and $R_{s|c}$.

Interpreting Cross-Ranks. The cross-ranks reveal how strongly concepts and skills co-occur for a given benchmark:

- If $R_{c|s}$ is *low* (i.e., skill neighbors have low concept ranks), this suggests that skill-similar samples are visually diverse, meaning the benchmark emphasizes general reasoning skills rather than specific visual domains.
- If $R_{c|s}$ is *high*, it indicates that skill-similar samples also share similar visual concepts, implying that the skills required are tied to particular domains.
- Conversely, a high $R_{s|c}$ indicates that concept-similar samples tend to involve similar skills, while a low $R_{s|c}$ indicates that concept neighbors are heterogeneous in skill demands.



(a) 5% data selection on LLaVA-1.5 (b) 10% data selection on LLaVA-1.5

Figure 1: Scatter of mutual rank difference and performance difference when prioritizing concept vs skill neighbors. The correlation remains consistent across different selection ratios, suggesting being concept- or skill-focused is an intrinsic property of benchmarks.

Analysis By comparing these two cross-ranks, we can infer the dominant alignment factor for the benchmark. **General-skill benchmarks** tend to have lower $R_{c|s}$ but higher $R_{s|c}$, since general skills apply broadly across many concepts, but concept neighbors mostly correspond to simple questions requiring similar low-level skills. On the other hand, **specific-skill benchmarks** tend to have higher $R_{c|s}$ but lower $R_{s|c}$, since samples that are similar in skill often share overlapping visual structures, whereas concept neighbors may not exhibit the specialized skills needed. This asymmetric pattern allows us to classify a benchmark as being primarily *skill-driven* or *concept-driven* without explicitly running both targeted selection strategies.

Results Fig 1 shows a scatter plot of the mutual rank difference ($R_{s|c} - R_{c|s}$) on the x-axis and the performance difference of model prioritizing concept vs skill on the y-axis. Data points closer to the lower-left correspond to the skill-targeted benchmarks, while the upper-right correspond to the concept-targeted. The simple cross-ranking heuristic successfully predicted the preferred alignment type, enabling an automated and lightweight benchmark-aware instruction selection policy. We find that this predictive approach is more consistent than naively combining the two strategies.

5.5 Skill description qualitative study

To verify that the proposed skill extraction pipeline captures meaningful and non-trivial information, we examined skill descriptions from the SQA-I benchmark and compared them to those of their nearest neighbors in the LLaVA-1.5 training mixture in Table 5. The retrieved neighbors demonstrate that the skill-based representation aligns closely with the steps needed to solve each question, beyond recognizing the concept entity in the picture.

For example, skills such as "interpreting graph trends across months" and "understanding growth requirements of plants from visual cues" highlight the kinds of subtle, context-dependent details that are not apparent from the image alone. This sanity check suggests that the skill embeddings are effective at grouping instructions by the type of reasoning required, rather than by surface-level visual similarity.

Table 5: Skill description comparisons between samples from SQA and their corresponding nearest neighbors in the LLaVA-1.5 training mixture.

Samples from benchmark	Nearest neighbors in training set
To interpret the graph accurately and identify temperature trends across the months.	The ability to interpret and analyze text and numerical data related to temperature ranges.
One must analyze the beak shape of the birds to determine adaptations for cracking hard seeds.	One needs to observe details of the bird's beak shape, feeding behavior, and surrounding environment.
Identifying and comparing the number of pink balls in each solution.	One must identify and differentiate between various types of balls in the image.
The ability to compare the number of seedlings in different pots and analyze growth differences is required.	Observation, interpretation of visuals, and understanding of growth requirements for plants.
Identifying organisms based on scientific names and recognizing their taxonomic relationships within a specific context.	Identifying the animal's species and recognizing its characteristics for accurate scientific naming.

6 Conclusion

This work highlights that vision-language benchmarks are not monolithic but fall naturally along a concept–skill axis: some predominantly reward grounding in visual concepts, while others emphasize skill-based reasoning. We show that existing instruction selection methods, which largely rely on embeddings, tend to prioritize concepts. Recognizing this bias is crucial for understanding why models succeed on certain tasks and underperform on others. To address this limitation, we developed a simple benchmark-aware instruction selection method that adapts training data to the dominant alignment factor of each benchmark, yielding consistent improvements across twelve diverse benchmarks. Beyond the empirical gains, our broader message is that advancing multimodal learning requires explicitly accounting for both what a model needs to recognize and the reasoning skills it must apply. We hope this benchmark categorization framework and our results will guide future research in model design, evaluation, and data selection, providing a clearer lens on the dual demands of concepts and skills in vision-language learning.

Limitations Our study has several limitations. The concept and skill taxonomy was derived semi-automatically and may lack nuance. Our selection method also presumes prior access to benchmark information, which may not be realistic for novel tasks. Furthermore, we evaluate tasks in isolation, without exploring multi-task generalization or interference effects. Lastly, the transferability of our findings across different model scales and architectures requires further investigation.

Ethics Statement This work does not raise any foreseeable ethical concerns. All datasets used are publicly available and intended for research purposes. No human or animal subjects were involved in this study.

Reproducibility Statement We have made efforts to ensure that our results are reproducible. All model architectures, training procedures, training hyperparameters as well as detailed descriptions of datasets, data processing steps, and evaluation protocols are provided in the are described in the main text (Section 4.1 and Section 4.2) and Appendix. Our method is simple and straightforward to reproduce, and we plan to release the source code upon acceptance of the paper.

LLM Usage Statement LLMs are only utilized to polish the writing and check for grammatical errors.

REFERENCES

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. URL https://arxiv.org/abs/2308.01390.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025.

Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, pp. 698–706. PMLR, 2018.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024.

Jiankang Chen, Tianke Zhang, Changyi Liu, Haojie Ding, Yaya Shi, Feng Cheng, Huihui Xiao, Bin Wen, Fan Yang, Tingting Gao, et al. Taskgalaxy: Scaling multi-modal instruction fine-tuning with tens of thousands vision task types. *arXiv preprint arXiv:2502.09925*, 2025.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 291–300, 2004.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. Concept-skill transferability-based data selection for large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5060–5080, 2024.

Bo Li*, Peiyuan Zhang*, Kaichen Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024. URL https://github.com/EvolvingLMMs-Lab/lmms-eval.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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. URL https://arxiv.org/abs/2301.12597.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

OpenAI. OpenAI Platform — platform.openai.com. https://platform.openai.com/docs/models. [Accessed 27-07-2025].

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.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M Patel, and Shao-Yuan Lo. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14247–14256, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0ff38d72a2e0aa6dbe42de83a17b2223-Paper-Datasets_and_Benchmarks_Track.pdf.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering, 2021. URL https://arxiv.org/abs/2107.09106.

Xindi Wu, Mengzhou Xia, Rulin Shao, Zhiwei Deng, Pang Wei Koh, and Olga Russakovsky. Icons: Influence consensus for vision-language data selection. *arXiv preprint arXiv:2501.00654*, 2024.

Zhiyang Xu, Ying Shen, and Lifu Huang. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11445–11465, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.641. URL https://aclanthology.org/2023.acl-long.641/.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A LLM Prompts

A.1 Skill Description Extraction

```
Here is a list of questions about an image:
[QUESTION_1]
[QUESTION_2]
[QUESTION_3]
Don't answer the above questions directly. What visual skills are
required to answer these questions? Answer in one short sentence
with less than 20 words without any extra reasoning.
```

A.2 Semantic Jaccard Similarity Calculation

```
You are an expert evaluator of semantic similarity. Compare the
following two sentences that describe the visual skills required
to perform a specific visual task. Rate the semantic Jaccard
similarity of the visual skills on a scale from 0.0 to 1.0.

Rules:
- 1.0 means the two sentences have exact semantically-equivalent
visual skills (even if phrased differently).
- 0.0 means the two sentences have no overlapping
semantically-equivalent visual skills.
- Output in json format with signature {reasoning: REASONING,
score: SCORE}

Sentence A: {sentence1}
Sentence B: {sentence2}

Output:
```

B Evaluation Details

We follow the evaluation pipeline of COINCIDE (Lee et al., 2024) for the overlapped benchmarks. The only minor difference is that we adopted `gpt-4o-mini` for judging the results for LLaVA-Bench. We also calculate the accuracy of VizWiZ by excluding the unanswerable subset. We found that the performance of the unanswerable subset is better for the worse models, likely because the models output "unanswerable" as the response to every question. For the new benchmarks, we directly utilize the implementation provided from the LMMS-Eval library (Li* et al., 2024).

C Baseline Reproduction Details

We reproduced the results for both Coincide (Lee et al., 2024) and PreSel (Safaei et al., 2025). Reproducing Coincide is straightforward since they fully open-sourced the codebase, and their

selection method supports selecting an arbitrary number of samples. PreSel requires first training a reference model on a random 5% subset of the dataset and then performs selection relying on the signal provided by the reference model. It is unfair to compare with PreSel at 5% directly since it would just be 5% of random samples. Therefore, the 5% baseline of PreSel was implemented by selecting an additional 5% of data with the reference model, effectively utilizing 10% of the data. However, PreSel fails to outperform other methods even with this advantage.

D Qualitative Study

D.1 Skill Description Comparison

Table 6 compare the visual skill descriptions of random samples from OK-VQA (skill-focused benchmark) and the skill description of their nearest neighbor in the LLaVA-665k training dataset. The skills required for the nearest neighbor closely resembles the skills needed for the benchmark sample, even if the visual subject does not exactly match.

Table 6: Skill description comparisons between samples from OK-VQA and their corresponding nearest neighbors in the LLaVA-1.5 training mixture.

Samples from benchmark	Nearest neighbors in training set
One must recognize the vehicle type and its design characteristics to determine its historical context and invention date.	Identifying the vehicle type and recognizing its historical context.
One needs to identify the meal's ingredients, presentation style, and cultural context to suggest a suitable side dish.	One must identify cultural elements in the dish's presentation, ingredients, and style.
Identifying the pastry type, size, and any visible toppings or fillings.	One must identify colors, textures, and shapes of the filling in the pastry.
One must identify size, shape, and features typical of buses versus vans to answer the question.	You need to identify and differentiate types of buses based on visual characteristics.
Identifying logos, labels, and packaging design details in the image to determine the origin of the beverage.	The ability to identify brand logos and labels on beverage packaging.

D.2 Semantic Jaccard Similarity Example

In this section we present an instance of semantic Jaccard similarity prediction which compares the overlap of visual skills between two skill descriptions. We adopt gpt-4o-mini as the LLM judge to generate the reasoning and the score of the Jaccard similarity. The two skill descriptions were independently extracted from the same instruction. The skill descriptions exhibit high consistency, suggesting that LLMs can serve as reliable and effective skill feature extractors.

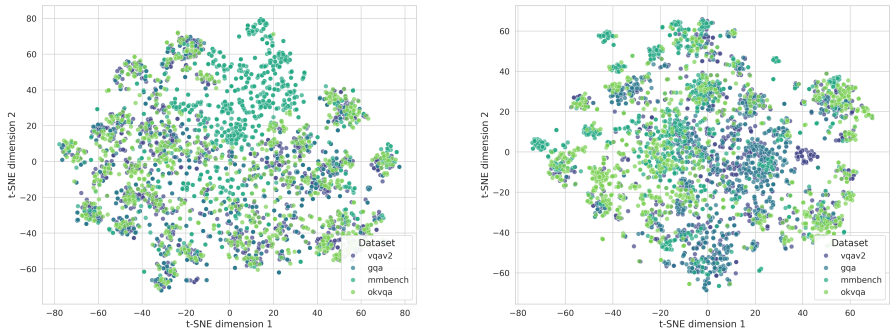
```
Sentence A: Visual skills required include the ability to
interpret visual layouts and identify UI elements in an app
interface.
Sentence B: Visual interpretation of app interfaces and
understanding visual design elements for navigation and usability.

Reasoning:  Both sentences discuss visual skills related to app
interfaces, specifically focusing on interpreting visual layouts
and identifying UI elements.  Sentence A emphasizes the ability
to interpret layouts and identify UI elements, while Sentence
B highlights visual interpretation and understanding design
elements.  The core concepts of interpreting and understanding
visual aspects of app interfaces overlap significantly, indicating
a high degree of semantic similarity.
Score:  0.8
```

D.3 Dataset Visualization in Concept and Skill Space

The dichotomy of concept and skill-focused tasks is motivated by the type of neighboring instructions that best improve its performance. Does the dichotomy also encode semantic

distinctions in its representation space? To this end, we visualize the projected representation of 2 concept-focused and 2 skill-focused datasets on both concept and skill space. Figure 2 presents the concept space visualization on the left and skill space on the right. Skill-focused datasets (e.g. OK-VQA) is more uniformly distributed in the concept space with other datasets while being more isolated in the skill space. This validates the mutual rank intuition where skill-focused tasks involve typical visual concepts but demand unique visual skills.



(a) Visualization in concept space (b) Visualization in skill space

Figure 2: Dataset visualization in concept and skill spaces. We observe that skill-focused tasks (e.g. OK-VQA) are more densely concentrated in the skill space and scattered in concept space.

D.4 Comparison Between Concept and Skill Neighbors

The goal of this section is to examine how concept and skill neighbors might differ. Figure 3 compares the nearest concept and skill neighbor retrieved from the same instruction randomly sampled from OK-VQA, a skill-focused task. We observe that although the retrieved concept nearest neighbor happens to be the exact matching image, the associated skill is different from the ones needed for the evaluation sample. On the other hand, the skill neighbors convey significantly more relevant skills and thus explains why OK-VQA is a skill-focused benchmark.

E Ablation Study

E.1 Concept-Skill Mutual Ranking

Table 7 studies the dependency of mutual ranking on the number of samples used for estimation. We compare the mutual rank calculated with the full dataset and $n = 50, 100, 500$ samples. Results show that mutual rank can be estimated with good accuracy with only 50 samples. This significantly decreases the computation requirement and increases the applicability of the method. We can confirm whether a task is concept or skill-focused with very few samples and only calculate the relevant embeddings.

E.2 Skill Extraction Model

The success of the skill-focused data selection depends on the LLM adopted to extract skill descriptions from the instruction. In this section we ablate the LLM to verify the robustness of the method against model choice. Specifically, we extracted skill descriptions with the open-sourced LLaMA-3.3-70B evaluate them on different benchmarks.

Table 8 compares the performance of skill-focused selection with LLaMA-3.3-70B and gpt-4o-mini. Both models exhibit a similar trend relative to the concept-targeted selection across different tasks – skill-targeted selection consistently performs better on skill-focused tasks and worse on concept-focused ones. This validates that the method is applicable even for open-sourced LLMs and the mutual ranking hypothesis generalizes across skill extraction model choices.

	Image	Instruction
Original Sample from OK-VQA		What type of building is this?
Concept Nearest Neighbor in LLaVA-665K		Please provide the bounding box coordinate of the region this sentence describes: a bronze statue is in the center of a park. Please provide the bounding box coordinate of the region this sentence describes: people walking around the statue. Please provide a short description for this region: [0.47, 0.55, 0.53, 0.62].
Skill Nearest Neighbor in LLaVA-665K		What special characteristics does this structure exhibit?

	Image	Instruction
Original Sample from OK-VQA		How tall is the tree that these fruit grow on?
Concept Nearest Neighbor in LLaVA-665K		Do the ripe bananas look yellow and long? Does the ceiling have white color? Where is this?
Skill Nearest Neighbor in LLaVA-665K		How tall is the girl? Answer the question using a single word or phrase.

Figure 3: Compare nearest neighbors in concept and skill space of random samples from OK-VQA.

E.3 Semantic Consistency of Skill Descriptions

LLM generation involves sampling and is inherently non-deterministic. Similar to how LLM-as-a-judge requires multiple calibration runs to verify the stability of the judge scores, it is essential to examine the consistency of skill descriptions over multiple runs. Consistency can be evaluated from two perspectives: representation semantic consistency and functional consistency.

Representation semantic consistency is defined as the overlap between extracted visual skills between different runs. Jaccard similarity is typically employed to measure overlap between sets, which is the size of intersection divided by the union. We extend Jaccard similarity to measure skill description similarity by taking the semantic equivalence of visual skills into account. Specifically, we measure the Jaccard similarity of visual skills while considering some visual skills might be stated differently over runs. For example, "Object recognition and color classification" and "Recognizing objects and pattern matching" would result in a semantic Jaccard similarity of $\frac{2}{3}$ since "object recognition" is semantically equivalent to "recognizing objects". We operationalize the calculation by employing LLM-as-a-judge with gpt-4o-mini as the judge model. The prompt is given in Appendix A.2 and an example of the output is in Appendix D.2.

Table 7: Ablation study on estimating concept-skill mutual ranking difference by sub-sampling data. The ranking can be accurately estimated even with as few as 50 samples from the evaluation dataset.

	Full Dataset			N = 500			N = 100			N = 50		
Task	$R_{s\|c}$	$R_{c\|s}$	R_{diff}	$R_{s\|c}$	$R_{c\|s}$	R_{diff}	$R_{s\|c}$	$R_{c\|s}$	R_{diff}	$R_{s\|c}$	$R_{c\|s}$	R_{diff}
VizWiZ	0.56	0.55	0.01	0.56	0.55	0.01	0.55	0.53	0.02	0.54	0.54	0.01
LlaVa-Bench	0.57	0.61	-0.04	—	—	—	—	—	—	0.58	0.63	-0.05
VQAV2	0.61	0.68	-0.08	0.62	0.67	-0.05	0.61	0.70	-0.09	0.59	0.70	-0.10
TextVQA	0.67	0.51	0.15	0.65	0.50	0.15	0.65	0.50	0.15	0.74	0.52	0.23
GQA	0.58	0.63	-0.05	0.57	0.64	-0.06	0.56	0.59	-0.04	0.55	0.62	-0.06
MME	0.60	0.60	0.00	0.61	0.59	0.02	0.61	0.61	-0.01	0.60	0.61	-0.02
MMBench(en)	0.56	0.60	-0.04	0.56	0.59	-0.03	0.55	0.62	-0.07	0.62	0.65	-0.03
POPE	0.60	0.60	0.00	0.62	0.60	0.01	0.57	0.61	-0.04	0.57	0.63	-0.06
STVQA	0.57	0.58	-0.01	0.57	0.57	-0.01	0.56	0.60	-0.03	0.61	0.52	0.09
SQA-I	0.48	0.64	-0.16	0.48	0.62	-0.13	0.47	0.64	-0.17	0.48	0.74	-0.26
AI2D	0.47	0.65	-0.19	0.47	0.66	-0.19	0.43	0.64	-0.21	0.46	0.65	-0.19
OKVQA	0.59	0.73	-0.14	0.60	0.72	-0.13	0.61	0.75	-0.14	0.61	0.77	-0.16

Table 8: Ablation study on using alternative open source models (Llama-3.3-70b) to extract skill descriptions. Models trained on instructions selected with both models exhibit significant gains on skill-focused tasks, demonstrating robustness of the method on the model choice.

				gpt-4o-mini		Llama-3.3-70b	
Category	Task	Random	Concept	Skill	C-S	Skill	C-S
Concept	TextVQA	52.0	54.8	54.0	+0.8	54.0	+0.8
	GQA	52.7	54.0	53.5	+0.5	52.9	+1.1
	MME	1259.3	1296.4	1287.7	+8.7	1219.4	+77.0
Hybrid	LlaVA-Bench	66.7	66.3	67.5	-1.2	64.4	+1.9
Skill	SQA-I	65.9	65.7	67.6	-1.9	66.7	-1.0
	AI2D	50.8	49.2	53.0	-3.8	52.7	-3.5
	OK-VQA	45.2	43.2	48.0	-4.8	52.0	-8.8

We calculate the pairwise semantic Jaccard similarity across 3 runs of skill description extraction with gpt-4o-mini, averaging over samples and pairs. The average similarity is 0.54 which indicates a high semantic consistency, compared to the random baseline value of 0.16 (samples are randomly permuted).

Functional consistency is defined as the samples selected with skill descriptions over multiple runs leading to similar performance on downstream tasks. To this end, we calculate the average and standard deviation of the performance on all benchmarks across different runs in Table 1. The low standard deviation bounds the variance of skill descriptions, suggesting high functional consistency.

E.4 Manual Verification of Skill Description Soundness

Skill descriptions are automatically generated by LLMs. In order to verify whether the descriptions truly reflect the visual skills necessary for performing the instruction, we constructed a manual survey to collect human feedback. 50 instructions were randomly sampled from LLaVA-665k. For each instruction we select its corresponding skill description as the correct option and a random skill description from the dataset as the incorrect option to construct a binary classification task. We then ask a human annotator to select the correct skill description that is necessary to perform the instruction.

We recruited 9 human annotators with the randomized survey and calculated the accuracy of human labels. The average accuracy is 94% with a standard deviation of 0.13. The highly consistent human agreement verifies the soundness of the skill descriptions.