# Does VLA Even Know the Basics? Measuring Commonsense and World Knowledge Retention in Vision–Language–Action Models

**Anonymous ACL submission**


## Abstract

Embodied Vision–Language–Action (VLA) models are typically obtained by fine-tuning powerful pretrained VLMs on robotics data, yet it is unclear how much commonsense and factual knowledge they retain after adaptation. Failures on knowledge-sensitive tasks are ambiguous, conflating missing knowledge with poor generalization of low-level control. We introduce Act2Answer, a lightweight protocol that adapts VLM knowledge benchmarks to VLA evaluation by requiring agents to answer through action. Each question becomes a short tabletop episode where the agent performs a single object-placement action to select among candidate answers, yielding an action-grounded success rate with reduced control confounds. We curate a large suite of such environments across diverse commonsense and world-knowledge categories and introduce layerwise intent probing to localize answer-relevant information across the VLM backbone and action head. In a large-scale study of 9 VLA models and 9 VLM baselines, we systematically rank models across categories, finding that VLAs retain simple perceptual concepts but often lose richer semantic knowledge relative to their source VLMs, that VQA co-training improves knowledge retention, and that answer-relevant signals peak in middle VLA layers but attenuate upper.


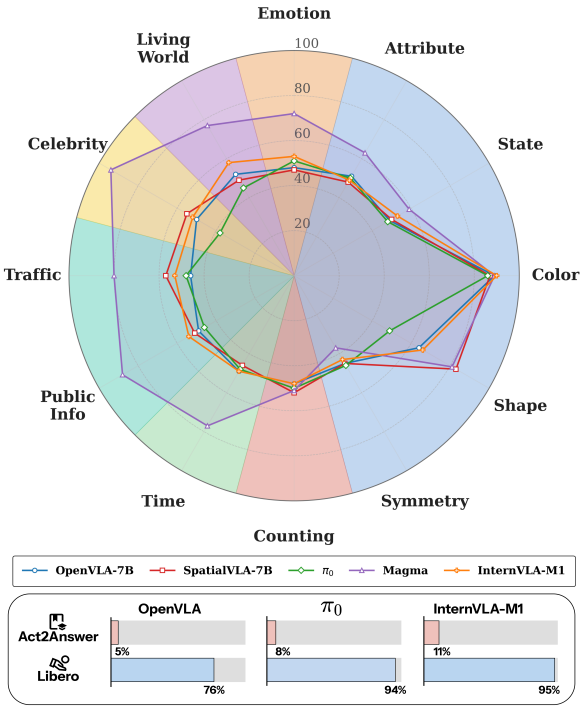


Figure 1: Knowledge evaluation results for seven state-of-the-art VLA models across diverse knowledge domains on the Act2Answer Task Suite. The bottom panel shows model performance averaged across all environments on both Act2Answer and LIBERO (Liu et al., 2023)

## 1 Introduction

Embodied agents are increasingly studied as candidates for deployment in everyday environments, such as households (Shukla et al., 2024; Mees et al., 2022; Liu et al., 2023) and retail (Soshin et al., 2025; Liu et al., 2025) settings. To operate effectively in such contexts, their actions must be grounded in a rich model of the world - what objects are, how they are typically used, and which behaviors are appropriate in a given situation - tightly linking low-level motor control with commonsense, world-level reasoning. Vision–Language–Action (VLA) models are widely proposed as the foundation for such agents and are often advertised as open-world-generalizable (Black et al., 2024; Kim et al., 2024; Yang et al., 2025; Qu et al., 2025), so we naturally expect them to handle previously unseen environments and objects with at least a basic level of appropriate generalization.

Yet, the rapidly expanding VLA literature has largely centered on manipulation-centric success. Current benchmarks (Liu et al., 2023; Mees et al., 2022; Shukla et al., 2024; Soshin et al., 2025; Li et al., 2024a; Zhang et al., 2024) ask whether agents

can complete complex tasks under perturbations, domain shifts, or new layouts, but they rarely ask whether agents retain even basic commonsense about objects, scenes, and goals after robotics training. Many VLA models are obtained by fine-tuning strong Vision-Language-Models (VLM) backbones on control tasks, and it is often tacitly assumed that the underlying world knowledge is preserved, even though we lack systematic ways to measure how much commonsense and factual knowledge is kept or catastrophically forgotten. In contrast, the VLM community has developed a rich ecosystem of benchmarks that explicitly probe world knowledge and commonsense. VLA evaluation, however, remains almost entirely task-success–centric: once a VLM backbone is fine-tuned into a policy that outputs actions, performance is usually reduced to success rates in manipulation or navigation domains, with little attention to whether the original knowledge is still present and accessible. As a result, **we currently lack a principled way to test what a VLA model still knows after robotics fine-tuning.**

We introduce ACT2ANSWER, an embodied evaluation protocol that adapts VLM knowledge benchmarks to VLA models by requiring action-based answer selection instead of text generation. Each question becomes a short simulated episode with a simple selection action, reducing confounds from long-horizon planning and low-level control. Prior action-based semantic evaluations (Zitkovich et al., 2023; Kim et al., 2024) are limited in scope; in contrast, ACT2ANSWER enables systematic benchmark adaptation across diverse commonsense and world-knowledge categories and supports layer-wise analyses that localize answer-relevant information within VLA models. Specifically, our key contributions are as follows:

1. We propose ACT2ANSWER, an embodied evaluation benchmark suite that adapts VLM knowledge tasks into action-based simulated episodes, making VLA knowledge measurable through behavior rather than text generation alone.

2. By adapting existing VLM benchmarks, we curate a large-scale and diverse set of environments for systematically evaluating commonsense and world knowledge in VLA models. In total, we collect 1720 unique binary questions across 12 categories, including attribute, state, color, symmetry, shape, emotion, celebrity, living world, counting, time, traffic, public info.

3. We present a large-scale empirical study of 9 modern VLA systems and 9 strong VLM baselines, systematically ranking models across knowledge categories. We find that current VLA models retain simple perceptual concepts but often lose richer semantic knowledge relative to their source VLMs, and that VQA co-training improves knowledge retention.

4. We introduce layerwise intent probing - linear classifiers trained on per-layer representations to predict the agent's option selection - to quantify how answer-relevant information is distributed across model depth. We show that relevant information can remain internally represented even when the model fails to translate it into the correct action.

## 2 Related Works

### 2.1 Evaluation of VLA Models

Current VLA benchmarks primarily evaluate manipulation success, emphasizing control generalization across tasks, scenes, embodiments, and language variations rather than explicit knowledge assessment. Benchmarks such as LIBERO (Liu et al., 2023), CALVIN (Mees et al., 2022), VLABench (Zhang et al., 2024), and BEHAVIOR-1K (Li et al., 2024a) measure whether agents can complete language-conditioned tasks, while MIKASA (Cherepanov et al., 2025), (Pugacheva et al., 2025), and VL-Think (Kachaev et al., 2025) probe memory, robustness, and transfer. However, across these settings, evaluation remains largely grounded in manipulation success and mostly tests only shallow semantics or primitive concepts, leaving commonsense and world knowledge largely unmeasured.

Recent works has increasingly considered knowledge transfer from VLMs to VLA models. A common evaluation strategy, used in works such as (Cai et al., 2026) and (Chen et al., 2025), is to test the VLM component on VQA-style benchmarks by decoding textual answers. This approach remains indirect: it measures whether the VLM component can still produce the correct textual answer, but not whether the VLA can use that knowledge when choosing an action. Consistent with this, Zhang et al. (2026) show that strong VQA performance

of VLM models does not necessarily translate into stronger VLA embodied behavior. In contrast, our Act2Answer protocol targets commonsense knowledge while preserving the embodied nature of evaluation: instead of answering in text, the agent must express its choice through action. This makes it possible to test whether knowledge remains usable in an action-conditioned setting, rather than only whether it can still be verbalized by the underlying VLM.

### 2.2 Knowledge and Commonsense in VLM

The VLM community has developed a broad set of benchmarks for evaluating multimodal understanding. Examples include GQA (Hudson and Manning, 2019) for compositional visual reasoning, TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2021) for question answering that requires reading text in images, AI2D (Kembhavi et al., 2016) for diagram understanding, ScienceQA (Lu et al., 2022) for multimodal science questions and MMMU (Yue et al., 2024) for large-scale, multidisciplinary multimodal evaluation. These benchmarks have become standard tools for assessing the knowledge and reasoning capabilities of VLMs. It is therefore essential to use such established benchmarks when evaluating knowledge in VLA systems. Our Act2Answer protocol enables this by converting suitable items into an embodied binary-decision format, making it possible to test knowledge in an action-based setting while retaining the benchmark grounding provided by existing VLM evaluations.

## 3 Methodology

### 3.1 Decomposing Embodied Task Success

Task success in embodied settings is not a single ability, but the outcome of several interacting factors. An agent must perceive the scene correctly, know which action is appropriate, execute that action reliably, and operate under the constraints of a particular environment. We use these four components: perception, knowledge, control, and environment - as a coarse conceptual decomposition of embodied task success. This decomposition matters because the same observed success rate may arise from very different underlying causes: strong knowledge but weak control, weak knowledge but strong motor routines, or even shortcut exploitation and favorable dynamics. As a result, end-to-end task success is often not diagnostic of what a VLA actually knows. Current benchmarks typically collapse these factors into a single outcome, making it difficult to tell whether failure reflects missing knowledge, perceptual error, motor difficulty, or environmental complexity. If the goal is to study knowledge in embodied agents, evaluation must therefore move beyond undifferentiated task success and more carefully isolate the contribution of knowledge from other sources of performance.

### 3.2 Commonsense Knowledge

What matters for embodied agents is not knowledge in the abstract, the question is what kinds of knowledge should actually be evaluated in embodied agents. Prior work such as Cosmos-Reason1 (Azzolini et al., 2025) has highlighted the importance of physically grounded reasoning over space, time, and causal constraints. However, real-world decision making often depends on a broader range of knowledge, including social roles, norms, quantities, biological constraints, and cultural context.

In this work, we use the term *commonsense knowledge* to refer to knowledge of this kind that can affect which action is appropriate in a given situation. Our goal is not to propose a universal taxonomy of embodied knowledge, but to introduce a practical set of knowledge categories that helps structure benchmark design and analysis. These categories are used to guide task selection, improve coverage across different types of commonsense knowledge, and support finer-grained error analysis. Because many embodied decisions are compositional, individual tasks may involve multiple factors at once. The proposed categories are therefore intended as a practical framework for organizing evaluation, rather than as a strict partition of embodied knowledge. Following this principle, we group the knowledge targeted in our benchmark into six broad domains:

**Physical** world knowledge covers properties of objects and materials, object states and state changes, object identity under occlusion or motion, spatial relations, affordances, intuitive mechanics, visibility and optics, thermodynamic processes, and physical causality or plausibility. These categories capture whether an agent can understand what objects are, what can be done with them, and what physical outcomes are possible.

**Temporal** knowledge covers action semantics, event segmentation, temporal order, duration, delayed effects, goal–subgoal structure, and short-horizon prediction or planning. These categories

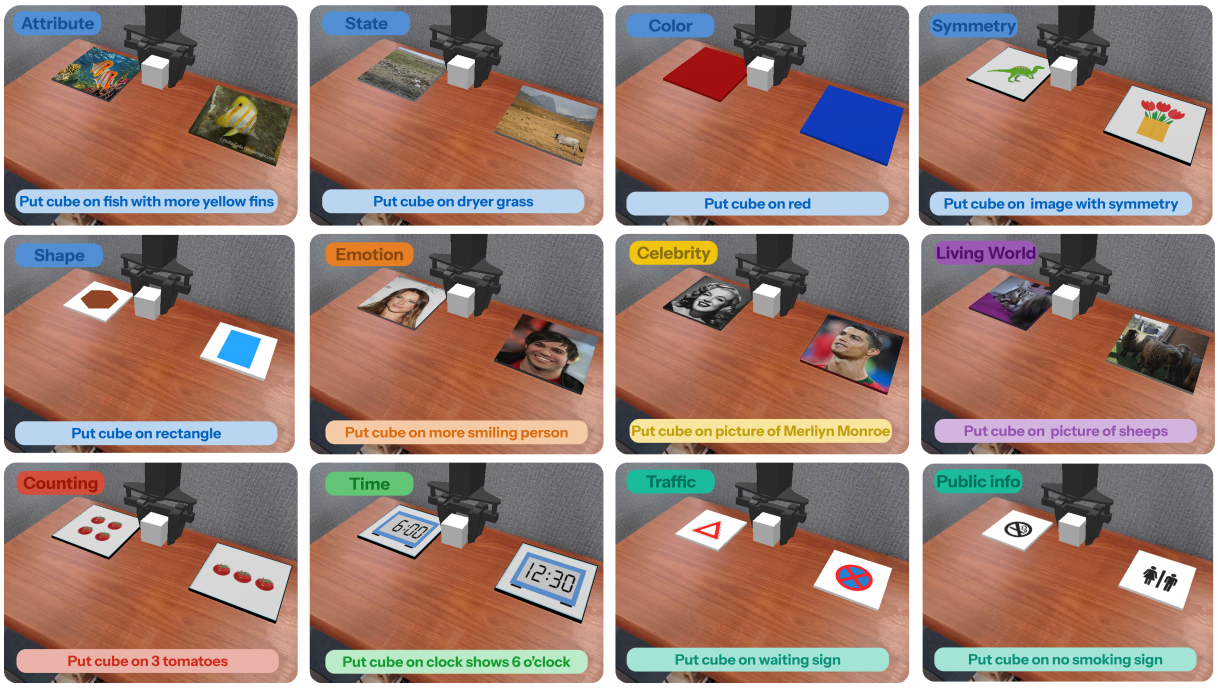


Figure 2: Act2Answer episodes examples for testing VLA models, built on top of VLM benchmark questions. In each episode, the embodied agent must interpret a natural-language instruction and control the robot arm to move the cube onto the correct answer plate.

capture whether an agent can interpret what is happening over time and choose an action that is appropriate not only for the current frame, but for the unfolding event.

**Quantitative** knowledge covers counting, magnitude comparison, measurement, rates or proportions, and basic resource accounting such as available space, remaining time, or sufficient quantity. Many embodied decisions are not purely semantic: they depend on whether there is enough room, enough material, or the right relative amount to make an action succeed safely and correctly.

**Biological** knowledge covers distinctions between living and non-living entities, bodily vulnerability, injury risk, food safety, allergy or toxicity, animal and plant needs, and age- or development-dependent constraints. Such knowledge is often necessary for acting appropriately around humans, animals, and biologically meaningful objects.

**Social** knowledge covers agent roles, emotions, intentions, beliefs, relationships, communication, joint attention, cooperation, and conflict. In embodied settings, correct action often depends not only on physical state, but also on who the agents are, what they know, what they want, and how they are interacting.

**Normative** knowledge covers moral and conventional norms, safety rules, institutional rules, role-based obligations, contextual appropriateness, customs, symbols, and culture-dependent interpretation. These categories matter whenever multiple physically possible actions exist, but only some of them are socially acceptable, safe, or contextually appropriate.

## 4 Act2Answer: Embodied Evaluation of Knowledge

The problem of knowledge evaluation in VLA models calls for a more direct way to assess *action-relevant* world knowledge. Current benchmarks often entangle knowledge with perception, control, and environmental complexity, while text-based evaluation of the VLM part does not reveal whether that knowledge can still guide action. The key unresolved issue is whether a model can not only possess relevant knowledge, but also act on it. To address this, we introduce ACT2ANSWER, a simple embodied evaluation protocol designed to test whether a VLA can express knowledge through action. A useful parallel comes from cognitive science, where knowledge in nonverbal agents is often evaluated through knowledge-conditioned actions rather than verbal reports. Inspired by this paradigms, ACT2ANSWER replaces textual answering with a minimal action that reveals the model's

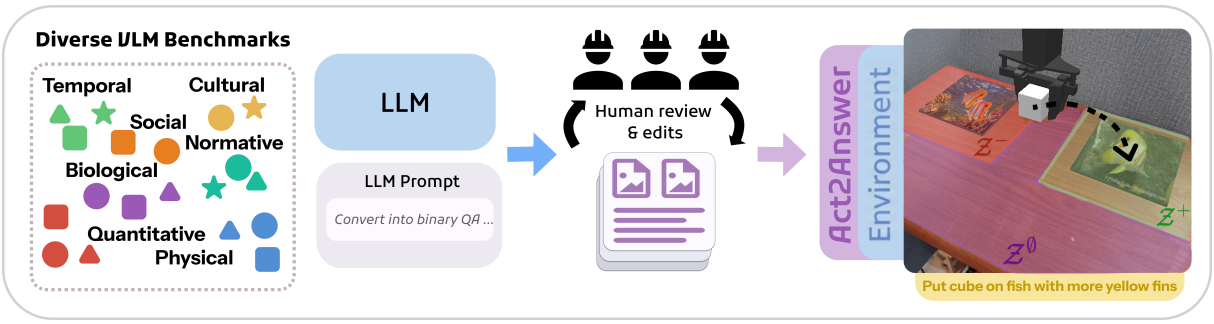


Figure 3: Overview of the data curation pipeline used to construct the Act2Answer task suite from VLM benchmarks, including selection, filtering and normalization, and conversion

choice. The goal is to evaluate whether an agent not only retains relevant world knowledge, but can also use it when selecting an action in a physically grounded setting.

### 4.1 Evaluation Protocol

The protocol is intentionally simple. Each episode is built from a VLM-style question together with one or more candidate images, such as a pair or a small grid, corresponding to possible answers. These images are placed at known positions in the scene, and the textual instruction is provided to the agent in the standard way. The agent must indicate its answer through a minimal embodied response by moving its end-effector and placing a cube on the selected image. The episode ends immediately after this answer action, and the prediction is scored as correct if the chosen image matches the correct answer. The protocol is designed to reduce the extent to which success is dominated by control difficulty or incidental environmental variation, and thereby make the outcome more directly informative about knowledge. More specifically, ACT2ANSWER is constructed so that success depends primarily on whether the agent can perceive, interpret the instruction, and use the required knowledge to select the correct answer, while minimizing the contribution of motor complexity and other low-level execution challenges.

| VLM Benchmark | Domain |
|---|---|
| MLLM-CompBench | Emotion, Attribute, State |
| IconQA | Time, Shape, Symmetry, Counting |
| MMBench | Celebrity |
| OK-VQA | Living World |
| VL-Think | Public Info, Traffic, Color |

Table 1: VLM benchmarks and knowledge domains used for adaptation to the Act2Answer protocol.

To this end, we deliberately reduce motor complexity: the required action is short-horizon, physically simple, and does not depend on difficult grasping or specialized manipulation skills. The resulting evaluation therefore does not fully isolate knowledge, but is intended to make the observed outcome more directly informative about whether the relevant knowledge remains available for action, rather than whether the agent can solve a difficult control problem.

### 4.2 Tasks Description

To evaluate VLA knowledge under the proposed knowledge categories, we construct the ACT2ANSWER task suite by adapting existing, community-established VLM benchmarks to an embodied setting rather than creating new question sets from scratch. This choice allows us to ground evaluation in widely used and already validated benchmark sources, while testing whether the same underlying knowledge remains available for action after adaptation from VLM to VLA. For each broad knowledge category, we select a representative subcategory, Table 1 summarizes the correspondence between the source VLM benchmarks and the knowledge categories used in our evaluation. Each benchmark item is then converted into a standardized ACT2ANSWER episode: the original question is reformulated as an instruction to the agent, and the candidate answers are instantiated as visual options placed in the scene. The agent must answer by performing the same minimal action defined in our protocol, namely selecting the correct option through object placement.

### 4.3 Data Curation

Starting from a diverse pool of existing VLM benchmarks, we first select tasks that match the selected knowledge categories. Because current

VLA models remain limited in long-context instruction following, we further filter the selected items by instruction length. We also apply image-level filtering: since many VLA models operate on relatively low visual resolution, human annotators remove examples in which the relevant objects are too small or visually ambiguous to be reliably perceived. After this selection stage, we unify the remaining tasks into a common embodied format. The original benchmark items may appear as open-ended questions or multiple-choice problems, so we convert them into binary decision tasks that are more suitable for action-based evaluation. To do so, we use an LLM to rewrite each example into a standardized two-option question while preserving its underlying knowledge requirement. Finally, these curated tasks are wrapped into an embodied environment built on the Simpler (Li et al., 2024b), where each instance can be executed under the ACT2ANSWER protocol. The result is a standardized task suite for evaluating whether VLA models retain and use Commonsense knowledge. In total, the suite covers 12 categories and contains 1,720 unique binary-choice items, corresponding to 3,440 evaluation episodes after including both original and swapped left/right configurations. Figure 3 illustrates the data curation pipeline.

### 4.4 Soft Success Rate

To support consistent interpretation of VLA performance in ACT2ANSWER, we define a simple interpretation methodology based on Soft Success Rate (SR). An episode is counted as successful if the agent places the cube on the correct answer tile within a tolerance region $\epsilon$ around it (Figure 3). To formally define this, let $p \in \mathbb{R}^2$ be the final 2D position of the cube at the end of an episode. We define a tolerance radius $\epsilon$ around the target image center $p^+$ and the incorrect option image center $p^-$ to partition the workspace $\mathcal{W}$ into target ($\mathcal{Z}^+$), incorrect ($\mathcal{Z}^-$), and out-of-bounds (OOB) ($\mathcal{Z}^\emptyset$) regions as follows:

$$\mathcal{Z}^+ = \{p \in \mathcal{W} : \|p - p^+\| \le \epsilon\}, \quad (1)$$

$$\mathcal{Z}^- = \{p \in \mathcal{W} : \|p - p^-\| \le \epsilon\}, \quad (2)$$

$$\mathcal{Z}^\emptyset = \mathcal{W} \setminus (\mathcal{Z}^+ \cup \mathcal{Z}^-). \quad (3)$$

Then the Soft Success Rate (SR) over $N$ binary-choice tasks is the empirical estimate of the probability of landing in the target region:

$$\mathrm{SR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(p^{(i)} \in \mathcal{Z}^+\right). \quad (4)$$

Based on the probability mass distribution among these regions, we interpret performance in three regimes:

1. **Instruction or perceptual failure (SR $< 0.5 - \Delta$):** The model fails to ground the visual options or the instruction. This manifests either as a dominant probability of out-of-bounds placement that dwarfs the valid regions ($\mathbb{P}(p \in \mathcal{Z}^\emptyset) \gg \max\{\mathbb{P}(p \in \mathcal{Z}^+), \mathbb{P}(p \in \mathcal{Z}^-)\}$), or as a systematic semantic misunderstanding where the model consistently favors the incorrect option ($\mathbb{P}(p \in \mathcal{Z}^-) \gg \mathbb{P}(p \in \mathcal{Z}^+)$). In both cases, the probability mass in the target zone drops significantly below random chance.

2. **No reliable usable knowledge ($|$SR $- 0.5| \le \Delta$):** The model correctly grounds the actionable regions ($\mathbb{P}(p \in \mathcal{Z}^\emptyset) \approx 0$) but lacks the specific knowledge required to select the correct option. This results in a near-uniform random guess or a severe positional bias between the two candidate regions, yielding $\mathbb{P}(p \in \mathcal{Z}^+) \approx \mathbb{P}(p \in \mathcal{Z}^-) \approx 0.5$ when evaluated across swapped spatial configurations.

3. **Evidence of usable knowledge (SR $> 0.5 + \Delta$):** The model successfully grounds the actionable regions and correctly leverages its internal semantic knowledge to skew the action distribution toward the target zone, such that $\mathbb{P}(p \in \mathcal{Z}^+) > \mathbb{P}(p \in \mathcal{Z}^-)$.

Finally, to reduce positional bias, we evaluate each example in both its original and swapped left/right versions and report the average score across the two. This makes the metric more robust to systematic side preferences.

### 4.5 Linear Intent Probing

Beyond task success, we measure whether answer-relevant information is present in a model's internal representations using *layerwise intent probing*. For each episode, we define an intent label $y \in \{0, 1\}$ indicating the selected option region (forced-choice answer). For a given VLA model, we extract hidden states from the last tokens of every transformer layer including both VLM and Action Expert parts. Since the action expert can be iterative (for example, 10 steps of flow-matching) we extract activations from every iteration. For each VLA we then train a linear probing independently at each layer

Table 2: Results of VLA and VLM models on emotion, attribute, state, and additional visual property tasks under the proposed Act2Answer evaluation protocol.

| Model | Social | Physical | | | | | Quantitative | Temporal | Normative | | Cultural | Biological |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Emotion | Attribute | State | Color | Shape | Symmetry | Counting | Time | Public Info | Traffic | Celebrity | Living World |
| InternVL3.5-8B | 95% | 68% | 64% | **100%** | 89% | 69% | 52% | 99% | 85% | 75% | 99% | 91% |
| InternVL3.5-38B | **99%** | **73%** | 68% | **100%** | 96% | **83%** | 59% | **100%** | **94%** | 81% | **100%** | 96% |
| Ovis2.5-9B | 89% | 69% | **69%** | **100%** | **98%** | **83%** | 59% | 99% | 88% | 85% | **100%** | **97%** |
| Qwen2.5-7B | 89% | 64% | 68% | **100%** | 90% | 78% | 62% | 99% | 80% | 86% | **100%** | 94% |
| Qwen2.5-32B | **99%** | 69% | **69%** | **100%** | 93% | **83%** | 61% | 99% | 85% | 86% | **100%** | 96% |
| Qwen3-8B | 86% | 68% | 67% | **100%** | 97% | 81% | **65%** | 98% | 83% | **93%** | **100%** | 95% |
| Qwen3-32B | 92% | 67% | 66% | **100%** | **98%** | **83%** | 59% | **100%** | 87% | 90% | **100%** | 86% |
| Prismatic-VLM-7B | 82% | 59% | 61% | 96% | 85% | 67% | 52% | 96% | 75% | 76% | 99% | 82% |
| Paligemma-3B | 53% | 50% | 52% | 47% | 48% | 49% | 49% | 51% | 48% | 49% | 48% | 49% |
| OpenVLA-7B | 48% | 51% | 49% | 89% | 64% | 45% | 48% | 49% | 49% | 46% | 50% | 52% |
| OpenVLA-7B SFT | 41% | 45% | 44% | 82% | 53% | 38% | 46% | 37% | 47% | 46% | 42% | 45% |
| OpenVLA-7B RL | 46% | 50% | 47% | 88% | 61% | 44% | 50% | 48% | 46% | 48% | 47% | 52% |
| SpatialVLA-7B | 47% | 48% | 50% | 87% | 83% | 45% | **52%** | 46% | 51% | 57% | 55% | 49% |
| $\pi_0$ | 51% | 50% | 48% | 86% | 49% | **46%** | 50% | 48% | 46% | 48% | 38% | 45% |
| Magma | **72%** | **63%** | **59%** | 89% | **81%** | 37% | 51% | **77%** | **88%** | **80%** | **94%** | **77%** |
| InternVLA-M1 | 53% | 49% | 53% | **90%** | 66% | 43% | 48% | 49% | 54% | 53% | 52% | 58% |

activations to predict $y$ and report probe accuracy as a function of layer number. For Action Expert layers we make this procedure independently for all iterations. We aggregate all linear probing classifiers with the final-layer accuracy $A_{\text{last}}$ and the best accuracy over layers $A_{\max}$. High $A_{\max}$ but low $A_{\text{last}}$ suggests that answer-relevant information is present in intermediate representations but is attenuated in later layers used for action prediction.

### 4.6 Evaluation and Results

Our goal is not to establish a new leaderboard, but to obtain an initial picture of what current VLA systems actually know under the Act2Answer protocol. Using Simpler (Li et al., 2024b), we evaluate several recent VLAs, including $\pi_0$ (Black et al., 2024), OpenVLA (Kim et al., 2024), MAGMA (Yang et al., 2025), and SpatialVLA (Qu et al., 2025), and compare them to strong VLM baselines such as Qwen2.5-VL, Ovis, PaliGemma, and InternVL (Table 2). Unless noted otherwise, we evaluate all VLA models using their original released checkpoints, without additional task-specific fine-tuning, the only exception is a separate set of supplementary ablations for OpenVLA. We then analyze the results to answer a set of research questions about what kinds of knowledge current VLAs preserve, lose, and remain able to use in action.

**RQ1: How well do current VLA models handle simple primitives?** We begin by testing whether VLA models can solve tasks built around basic perceptual concepts such as COLOR and SHAPE. These categories serve as a useful lower bound: if models fail here, it would suggest a broad inability to use even simple visual distinctions in action. In practice, all evaluated models perform strongly on such tasks, with high success rates across primitive categories. This indicates that simple Physicall knowledge is largely preserved and remains action-usable in current VLA systems.

**RQ2: Can VLA models handle more complex semantic concepts?** We next turn to categories that require richer semantic interpretation, such as ATTRIBUTE, STATE, and EMOTION. Unlike primitive visual distinctions, these tasks require the model to access more abstract knowledge in order to choose the correct answer. Here performance drops sharply: in ATTRIBUTE in particular, most models behave close to random chance, with similar trends in several other higher-level categories. This suggests that current VLAs struggle once correct action depends on more than shallow perceptual matching.

**RQ3: Is knowledge preserved when moving from VLMs to VLAs?** To estimate how much semantic capability survives the transition from VLM to VLA, we compare each VLA to its original VLM checkpoint in an action-free probing setup. Given the first frame, we ask the VLM: *"Do you see the <board_name>? Answer yes or no. If yes, specify where: left, center, or right."* A prediction is counted as correct only if both the board identity and its position match the ground truth, yielding a rough measure of semantic grounding that can be compared to Act2Answer performance. The results show a large gap: across most domains, the original VLMs outperform their VLA counterparts by roughly 20–40 points. This points to substantial knowledge degradation during adaptation from vision-language pretraining to embodied policy learning.

**RQ4: Where does the knowledge go?** A natural follow-up is whether this knowledge is truly erased or whether it remains somewhere in the hidden representations but is no longer accessible for action. To study this, we perform linear probing over all layers of VLA models on categories such as ATTRIBUTE, STATE, EMOTION, and COUNTING. The results in Figure 4 reveal a consistent pattern: intermediate layers of the VLM backbone are often above chance, suggesting that task-relevant information is still present, but performance declines toward the final layers used for action prediction, often approaching random guessing. This suggests a bottleneck between semantic representation and action generation: the model may still contain the relevant knowledge internally, but fails to decode it into the correct action.

**RQ5: Does co-training on vision-language data help preserve knowledge?** We also ask whether co-training on image-text data alongside robot data improves knowledge retention. To test this, we include additional VLA baselines trained with joint vision-language and robotics supervision. The overall trend is positive: such models perform better on average across most categories than their purely robotics-trained counterparts. While the gains are not uniform in every domain, the results suggest that continued visual-language supervision helps preserve action-relevant semantics during embodied training.

**RQ6: How does downstream fine-tuning affect knowledge?** Finally, we study whether additional downstream fine-tuning improves or harms knowledge retention. Using OpenVLA as a case study, we fine-tune the model with both SFT and RL on a small pick-and-place dataset that is not directly drawn from our benchmark, but does include visual and semantic perturbations. The results do not show consistent improvements. On the contrary, some categories, including STATE and COLOR, exhibit noticeable drops after fine-tuning. This suggests that standard downstream adaptation may further bias the model toward task-specific action optimization, sometimes at the expense of general semantic knowledge.

## 5 Conclusion

We introduced *Act2Answer*, a simple protocol for measuring basic knowledge in VLA models. In our setup, a VLA operates in a table-top environment with several candidate images and a natural-language prompt, and must answer by performing a minimal embodied action, such as placing a dummy object onto the image it believes to be correct. This design keeps interaction embodied, but makes the choice itself as close as possible to the multiple-choice formats used in VLM benchmarks, so that errors can be interpreted primarily in terms of (missing) knowledge rather than low-level motor difficulty.

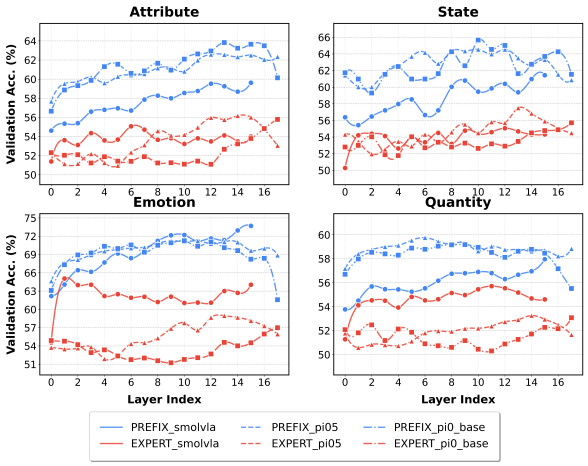


Figure 4: Probing results for internal representations of VLA models on four tasks from the Act2Answer task suite. In the legend, PREFIX labels indicate representations from the VLM component, whereas EXPERT labels indicate representations from the Action Expert component.

Our investigation reveals a striking "semantic tax" inherent in current VLA training paradigms. The transition from VLM to VLA tends to preserve low-level visual discrimination (e.g., color, shape, coarse object identity), yet it appears to severely erode higher-level conceptual understanding. The sharp drop in performance on Emotion, Attribute, and State tasks suggests that, in teaching models how to move, we are causing them to forget why they might need to move in the first place. A robot that can reliably grasp a cup but cannot distinguish a "dirty" cup from a "clean" one, or a "sad" human from a "neutral" one, is fundamentally limited in its usefulness as an assistant in everyday environments. These results indicate that simply fine-tuning VLMs on action data is insufficient: the next generation of embodied agents will require architectures and training objectives that actively preserve and align the backbone's semantic world model with its learned motor policies, rather than trading one for the other.

## 6 Limitations

Act2Answer is intentionally simple and efficient, but it also has important limitations. First, it measures whether knowledge appears to be preserved or lost, but it *does not explain why or where particular concepts are forgotten or retained*. Closing this gap *calls for complementary analyses*: probing intermediate representations, tracking how concept classifiers evolve layer by layer during fine-tuning, and relating specific training data or stages to specific patterns of forgetting and retention.

Second, our current tasks and question formats remain relatively simplified, so the measured "basic knowledge" may overestimate performance in realistic, high-stakes deployments. We envision richer benchmarks that keep the Act2Answer spirit but move closer to real-world conditions: for example, multi-step scenarios where the agent must choose the right tool and then execute a short manipulation, social situations where it must select an appropriate response to an expressed emotion, or physically grounded scenes where stability, containment, or support relations must be judged from cluttered views. Such benchmarks could be instantiated both in higher-fidelity simulators and in controlled real-robot setups, providing a more stringent test of commonsense under realistic sensory noise and embodiment constraints.

## 7 Ethical Considerations

Our work proposes an evaluation protocol and does not collect new human-subject data. All tasks are derived from publicly available VLM benchmarks under their original usage terms. Some categories involve images of people and public figures (e.g., Celebrity, Emotion), which are used solely for diagnostic evaluation of model knowledge and not to support surveillance or high-stakes inference about individuals. We view making knowledge degradation in VLA models measurable and visible as an important step toward safer and more reliable embodied AI systems.

We therefore invite the community to build on our initial step: to design more rigorous, realistic embodied commonsense benchmarks, and to develop VLA models and training paradigms that not only learn how to act, but also reliably retain and use what they know about the world.

## References


Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, and 1 others. 2025. pi*0.6: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*.

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, and 1 others. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, and 5 others. 2024. $\pi_0$: A vision-language-action flow model for general robot control. *Preprint*, arXiv:2410.24164.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, and 1 others. 2026. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*.

Xinyi Chen, Yilun Chen, Yanwei Fu, Ning Gao, Jiaya Jia, Weiyang Jin, Hao Li, Yao Mu, Jiangmiao Pang, Yu Qiao, and 1 others. 2025. Internvla-m1: A spatially guided vision-language-action framework for generalist robot policy. *arXiv preprint arXiv:2510.13778*.

Egor Cherepanov, Nikita Kachaev, Alexey K. Kovalev, and Aleksandr I. Panov. 2025. Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning. *Preprint*, arXiv:2502.10550.

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

Nikita Kachaev, Mikhail Kolosov, Daniil Zelezetsky, Alexey K. Kovalev, and Aleksandr I. Panov. 2025. Don't blind your vla: Aligning visual representations for ood generalization. *Preprint*, arXiv:2510.25616.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. 2024. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 28798–28827. Curran Associates, Inc.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. Openvla: An open-source vision-language-action model. *Preprint*, arXiv:2406.09246.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, and 16 others. 2024a. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *Preprint*, arXiv:2403.09227.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. 2024b. Evaluating real-world robot manipulation policies in simulation. *Preprint*, arXiv:2405.05941.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Preprint*, arXiv:2306.03310.

Weiheng Liu, Yuxuan Wan, Jilong Wang, Yuxuan Kuang, Xuesong Shi, Haoran Li, Dongbin Zhao, Zhizheng Zhang, and He Wang. 2025. Fetchbot: Object fetching in cluttered shelves via zero-shot sim2real. *arXiv preprint arXiv:2502.17894*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Advances in Neural Information Processing Systems*. Datasets and Benchmarks Track.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *Preprint*, arXiv:2112.03227.

Daria Pugacheva, Andrey Moskalenko, Denis Shepelev, Andrey Kuznetsov, Vlad Shakhuro, and Elena Tutubalina. 2025. Bring the apple, not the sofa: Impact of irrelevant context in embodied ai commands on vla models. *Preprint*, arXiv:2510.07067.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. 2025. Spatialvla: Exploring spatial representations for visual-language-action model. *Preprint*, arXiv:2501.15830.

Arth Shukla, Stone Tao, and Hao Su. 2024. Maniskillhab: A benchmark for low-level manipulation in home rearrangement tasks. *arXiv preprint arXiv:2412.13211*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Konstantin Soshin, Alexander Krapukhin, Andrei Spiridonov, Denis Shepelev, Gregorii Bukhtuev, Andrey Kuznetsov, and Vlad Shakhuro. 2025. Robobenchmart: Benchmarking robots in retail environment. *arXiv preprint arXiv:2511.10276*.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, Lars Liden, and Jianfeng Gao. 2025. Magma: A foundation model for multimodal ai agents. *Preprint*, arXiv:2502.13130.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Jianke Zhang, Xiaoyu Chen, Qiuyue Wang, Mingsheng Li, Yanjiang Guo, Yucheng Hu, Jiajun Zhang, Shuai Bai, Junyang Lin, and Jianyu Chen. 2026. Vlm4vla: Revisiting vision-language-models in vision-language-action models. *arXiv preprint arXiv:2601.03309*.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. *Preprint*, arXiv:2412.18194.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

# A Prompt Template for VLM Evaluation

We use a standardized prompt template to evaluate the vision-language model (VLM) in a constrained multiple-choice setting. The prompt consists of three components: (i) an image placeholder, (ii) a natural language question, and (iii) a set of answer options.

**Prompt Template.**

```
<image> Do you see the {object}? If yes, specify
where: 'left' or 'right'.
A left
B right
Answer with the option's letter from the given
choices directly.
```

**Answer Format.** The model is instructed to output only the corresponding option letter (e.g., "A" or "B"), which ensures consistent and easily comparable predictions.

This transformation maps chance-level performance to 0 and perfect accuracy to 100, making it easier to compare models on a common scale.

# B Benchmark Sources and Data Construction

This appendix provides additional details on the source benchmarks used to construct the ACT2ANSWER task suite, namely MLLM-CompBench (Kil et al., 2024), IconQA (Lu et al., 2021), MMBench (Liu et al., 2024), OK-VQA (Marino et al., 2019), and VL-Think (Kachaev et al., 2025). Our benchmark selection followed a simple principle: we prioritized source tasks whose original evaluation format could be transferred into an embodied binary-decision setup with minimal change to the underlying knowledge requirement. In practice, this favored benchmarks with paired-image comparison, multi-image choice, or visually grounded multiple-choice structure. We further restricted the final pool to examples whose relevant visual evidence remains perceivable under the lower effective resolution typical of current VLA systems and whose answer can be expressed through a short action-conditioned instruction.

## B.1 Near Format-Preserving Transfers

*MLLM-CompBench.* MLLM-CompBench was the cleanest source benchmark for ACT2ANSWER. Its native format already consists of two candidate images and a comparative question, making it almost perfectly aligned with our embodied answer-selection setup. We therefore used it as the primary source for EMOTION, ATTRIBUTE, and STATE. Adaptation in this case was close to format-preserving: we retained the original image pair and only rewrote the question into a short imperative instruction suitable for VLA models. Because almost no additional restructuring was required, these categories provide one of the most direct comparisons between VLM and VLA behavior in our study.

*VL-Think.* VL-Think was also a natural fit for our protocol because its task structure is already close to an embodied semantic selection problem. We used it for PUBLIC INFO, TRAFFIC, and COLOR, where the relevant concepts are encoded in compact visual symbols or simple public conventions. Since the source benchmark already operates in a board-selection style setting, adaptation mainly consisted of unifying the instruction style and episode format with the rest of the ACT2ANSWER suite. Because VL-Think is specified in terms of small closed concept vocabularies rather than large candidate image pools, the corresponding 'Initial pool' entries in Table 3 report the number of source concepts rather than the number of raw image examples.

## B.2 Moderately Curated Transfers

*IconQA.* IconQA was used for TIME, SHAPE, SYMMETRY, and COUNTING. We retained examples whose answer could be directly instantiated as visual option selection in our embodied setup. Because IconQA uses schematic icon diagrams rather than cluttered natural images, it remains relatively robust under the lower visual resolution of current VLA systems. The candidate pools differ substantially in size because we selected only those subsets that were most compatible with VLA evaluation. SHAPE was drawn from a broader geometry pool, whereas SYMMETRY was treated as a narrower geometry subcategory, selected separately because it provided a cleaner match to our broader PERSPECTIVE domain and was qualitatively distinct from the remaining geometry-style questions.

For TIME, we retained only the more visually grounded clock-style subset that could be cleanly expressed as short embodied instructions. Overall, these choices were made to favor examples with clear visual evidence and a natural fit to the ACT2ANSWER setting. When an original item contained more than two answer candidates, we converted it into multiple binary-choice episodes by pairing the correct option with one distractor at a time.

| Source benchmark | Domain | Native format | Transfer type | Initial pool | Final eval |
|---|---|---|---|---|---|
| MLLM-CompBench | Emotion | image pair + comparative question | near | 5254 | 300 |
| | Attribute | | | 5386 | 300 |
| | State | | | 1066 | 300 |
| IconQA | Time | multi-image choice | moderate | 508 | 300 |
| | Shape | | | 12798 | 300 |
| | Symmetry | | | 362 | 300 |
| | Counting | | | 5001 | 300 |
| MMBench | Celebrity | multiple choice | moderate | 316 | 140 |
| OK-VQA | Living World | open-ended VQA | heavy | 2336 | 300 |
| VL-Think | Public Info | embodied board selection | near | 14 concepts | 300 |
| | Traffic | | | 24 concepts | 300 |
| | Color | | | 11 concepts | 300 |

Table 3: Per-category curation statistics for the ACT2ANSWER suite. 'Final eval' reports the total number of evaluation episodes after including both the original and swapped left/right configurations for each selected item. For VL-Think, the 'Initial pool' column reports the number of source concepts rather than the number of candidate image examples.

*MMBench.* We used MMBench as the source for the CELEBRITY category. We first identified candidate slices through benchmark metadata and question templates indicating person-recognition behavior, especially prompts centered on identity recognition. These source examples were then converted into binary image pairs suitable for embodied evaluation. The final celebrity subset was not transferred verbatim: we additionally re-curated the source pool by replacing several original identities with more broadly recognizable public figures. This was done to better target shared world knowledge rather than long-tail memorization of obscure identities. The resulting subset should therefore be viewed as a curated derivative of MMBench rather than a direct extraction.

### B.3 Heavily Curated Transfer

*OK-VQA.* OK-VQA was used for the LIVING WORLD category, covering animal identity, flora identity, and living-versus-nonliving distinctions. Because OK-VQA is natively open-ended and text-answer based, it required substantially heavier curation than the benchmarks above. We merged the official question and annotation files, computed majority answers and agreement statistics, and retained only examples with short, visually grounded answers and sufficiently stable annotator agreement. We then filtered the pool by question type, excluding prompts that were unlikely to transfer cleanly into a binary image-choice setup.

From this filtered pool, we extracted narrower subsets using question-template matching and answer-level filtering. We then constructed binary visual episodes by pairing correct images with hard negatives drawn from semantically plausible candidate pools. Since automatic pairing alone was not sufficient, we performed manual review to remove ambiguous cases and refine the final embodied instructions. When the validation split did not provide enough high-quality examples for a target subset, we supplemented it with carefully filtered train-split items. As a result, the OK-VQA-derived portion of ACT2ANSWER is best understood as a semi-automatically curated benchmark specifically optimized for embodied knowledge evaluation.

### B.4 Instruction Rewriting and Binary Conversion

A central design goal of ACT2ANSWER is to preserve the original knowledge requirement of the source benchmark while changing only the response modality from text-based answering to action-based selection. To achieve this, we used an LLM as a first-pass rewriting tool to convert source benchmark questions into short imperative instructions suitable for VLA models, and then applied human review and manual edits to refine the resulting prompts. Whenever possible, this rewriting remained close to surface rephrasing rather than semantic reformulation. For example, comparative prompts from MLLM-CompBench were converted into instructions of the form *"Put the cube on . . . "* while preserving the same comparison target and

candidate images.

Representative examples of the final instruction style include prompts such as *"Put the cube on the more smiling person"* for EMOTION, *"Put the cube on dryer grass"* for STATE, *"Put the cube on rectangle"* for SHAPE, *"Put the cube on the clock showing 6 o'clock"* for TIME, *"Put the cube on the no smoking sign"* for PUBLIC INFO, and *"Put the cube on the picture of sheep"* for LIVING WORLD. These examples illustrate the general design principle used throughout the suite: the final VLA-facing instruction is kept short, visually grounded, and as close as possible to the original semantic target of the source benchmark.

### B.5 Discussion

**How to prevent the erosion of world knowledge?** Finally, it is not yet clear how to systematically *prevent the erosion of world knowledge during VLA training*. Developing effective training procedures that retain the backbone's world knowledge while acquiring strong control policies remains an open challenge. Promising directions include multi-task and continual-learning schemes that interleave action prediction with general question answering (Zitkovich et al., 2023; Amin et al., 2025), regularization or distillation techniques that explicitly preserve VLM representations, and architectural decoupling of knowledge and control components. However, large-scale multi-task or continual training that keeps many skills active can be prohibitively compute-intensive for real-world systems, especially when combined with high-resolution perception and long-horizon control. This makes *it essential to explore more efficient paradigms and lightweight mechanisms to prevent forgetting* - such as parameter-efficient fine-tuning, targeted regularization, selective rehearsal, or sparsely updated knowledge modules - so that VLAs can become more capable without repeatedly sacrificing the commonsense they started with.

## C Details of Score Averaging

For LIBERO, all values for the compared VLA models were taken directly from the original papers. For Act2Answer, we used the results from Table 2. For each model, we first averaged accuracy across all environments, and then normalized this mean relative to random guessing (50%) using a linear rescaling:

$$\text{Normalized Score} = \frac{\text{Mean Accuracy} - 50}{50} \times 100. \tag{5}$$