AVA-Bench: Atomic Visual Ability Benchmark for Vision Foundation Models

Supplementary Material

Appendix

Disclosure of LLM Usage. Portions of this manuscript were polished for clarity and readability using an LLM. The LLM was not used to generate research ideas, design experiments, analyze data, or draw conclusions. All scientific content, methods, and results are the authors' original work.

We provide details omitted in the main paper.

A. AVA-BENCH Details

A.1. Atomic Visual Abilities (AVAs)

As mentioned in Section 3.1, AVAs are elemental visual capabilities that can be combined to address more complex visual reasoning tasks. The definitions and representative questions for each AVA can be found in Table 1. Additional qualitative illustrations are provided in Figure 2 and Figure 4 (b).

The 14 AVAs selected for AVA-BENCH are grounded in a thorough literature analysis.

1. **Compositional text-to-image (T2I) benchmarks.** Studies on controllable generation motivate core visual primitives—number, colour, texture, object identity, spatial relations, and more—used to construct compositional prompts [37, 98]. These primitives form an initial pool of candidate abilities.
2. **VQA question analysis.** We employ GPT-4 to summarize the visual skills demanded by VQA questions in various commonly-used datasets (VQAv2 [27], RealWorldQA [100], GQA [1], etc.), thereby enriching the pool with abilities emphasized by real-world questions.

Intersecting these two sources yields a concise yet crucial set of AVAs. Moreover, we focus strictly on pure perceptual tasks and exclude non-perceptual reasoning skills (e.g., historical context, mathematical reasoning). We provide more related work discussion in Appendix E.

A.2. Dataset Curation

Spatial Reasoning [95, 110]. We curate **11.5K** image pairs from **NYU-Depth V2** [80] (indoor scenes) and **LVIS** [30] and **Objects365** [78] (open-domain scenes), all containing instance segmentation annotations. In each image, two distinct, non-overlapping objects are selected—one highlighted with a blue bounding box (reference object) and another with a red bounding box (target object). The model must identify the relative spatial position of the red box with respect to the blue box, choosing from four multiple-choice options: *Left above*, *Left below*, *Right above*, and *Right below* (Figure 9). The preprocessing steps for dataset creation are summarized below:

- To prevent ambiguity in interpretation, we restrict each image to contain only one instance of the target and reference objects.
- Object pairs whose bounding boxes overlapped either horizontally or vertically were excluded, ensuring unambiguous assignment to the four spatial categories.
- Extremely small bounding boxes complicating localization were filtered out. Specifically, object instances covering at least 2% of the image area for NYU-Depth V2, and at least 0.2% for LVIS and Objects365, were retained.
- For every question, we have a target and a reference object with a spatial position (the relative position of the target based on the reference object). Each object class was ensured to appear in multiple spatial positions, with 40 samples per spatial position category. For each target object class in each spatial position, we ensure diversity of reference by selecting samples from 8 distinct reference classes and drawing 5 rows per reference, thereby preventing overfitting to a small set of co-occurring anchors.
- Each object class was ensured to appear in multiple spatial positions, with 40 samples per spatial position category. This prevents models from memorizing fixed layouts.
- For each target object class, reference object class, and spatial position category, an 80% train and 20% test split was ensured, for uniform distribution and fair evaluation.
- The question for each pair: *"Considering the relative positions of two objects in the image, where is the microphone (annotated by the red box) located with respect to the speaker (annotated by the blue box)? Choose from A. Left above, B. Left below, C. Right above, D. Right below."*

Counting [47, 105]. We curate a total of **13.6K** images from five datasets— **VQAv2** [27], **FSC-147** [74], **CARPK** [34], **LVIS** [30] , and **CrowdHuman** [49] —to evaluate object counting abilities across diverse domains, including open-domain scenes, natural objects, structured environments, and densely crowded contexts. Each im-

Atomic Visual Ability	Definition	Example Question
Counting	Determining the number of instances of an object	How many apples are in the image?
Localization	Identifying the location of an object in the image	Provide bounding box coordinate for bicycle.
Fine-Grained	Differentiating between similar sub-categories of objects	What species of fungi is in the image?
OCR	Reading and interpreting text visible in the image	What is written in the red bounding box in the image?
Absolute Depth	Estimating how far an object is from the camera	From the camera's perspective, estimate how far the closest point of the car (red box) is from the camera in real-world distance, in meters.
Relative Depth	Comparing distances of two objects from the camera	Which object is closer to the camera, the car (red box) or the cyclist (blue box) to the camera?
Orientation	Determining the facing direction or angle of an object	What is the orientation of the toy bus in the image?
Spatial	Inferring layout and spatial relations	Considering the relative positions of two objects in the image, where is the bicycle (red box) located with respect to the towel (blue box)?
Object Recognition	Identifying objects present in the image given bounding box	What is in the red bounding box in the image?
Scene Recognition	Identifying the broader environment or type of setting	What is the scene class of the image?
Action Recognition	Determining what action is being performed	Which action or activity is shown in the image?
Texture	Describing surface appearance or material of objects	What is the texture attribute of image?
Color	Identifying colors of objects given bounding box	What color is shown within the bounding box?
Emotion	Recognizing emotional expressions in humans given bounding box	What emotion is being shown in the image?

Table 1. **Atomic Visual Abilities (AVAs).** We identify 14 AVAs, serving as the foundational capabilities that can be combined to tackle complex visual reasoning tasks. For each AVA, we provide the definition and an example question in `ourbench`.

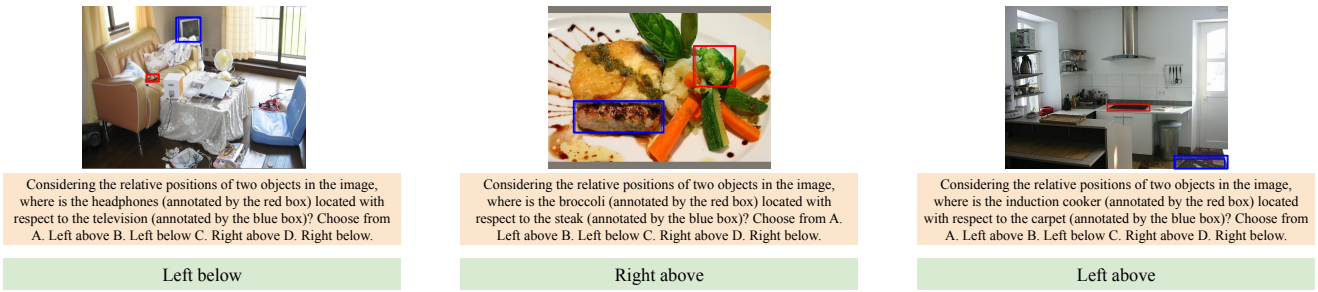


Figure 9. Examples of Spatial Reasoning AVA Samples.

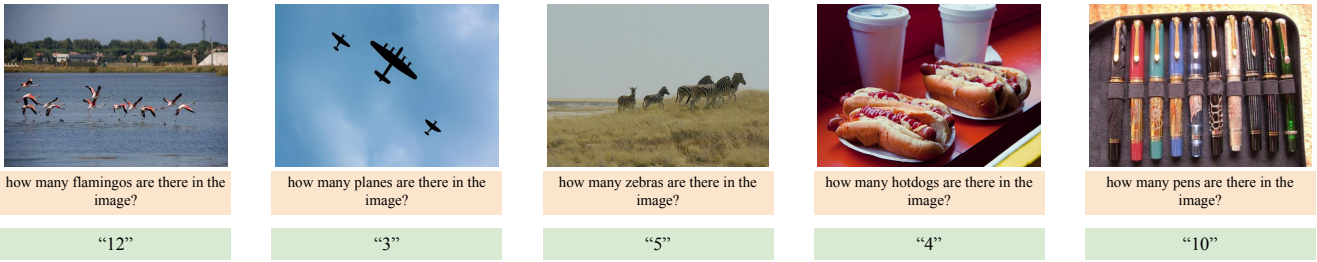


Figure 10. Examples of Counting AVA Samples.

age is paired with a question prompting the model to count the number of instances of a specified object category, and the model must return an integer-valued answer(Figure 10). The preprocessing steps for dataset creation are summarized below:

- To ensure valid supervision, we filter all samples to retain only those with non-zero object counts and with object_count ≤ 40.
- For each object category (object id), we require at least 4–5 distinct object count values to be represented, preventing overfitting to static object layouts.
- For each object count, we sample a fixed range of images per object_id—between 6 and 30 depending on the dataset—to balance frequency and diversity.
- Dataset-specific sampling rules are applied:
  - VQAv2 and LVIS: 15–30 images per object count; ≥ 5 count values per object id.
  - FSC-147: 6–12 images per object count; ≥ 4 count values per object id.
  - CARPK: 10–20 images per object count; ≥ 5 count values per object id.
  - CrowdHuman: object count capped at 40; 25–50 images per count level.
- An 80% train / 20% test split is maintained for each object id and count level to ensure balanced distribution during evaluation.
- The question for each image is: *"How many [object] are there in the image?"*, where [object] refers to the annotated target category.

Fine-grained [29, 71, 114]. We curate a total of **9K images** from five fine-grained recognition domains—**Bird**, **Animal**, **Fungi**, **Plant**, and **Object**—to assess species-level recognition capabilities. The dataset sources include CUB-200-2011 [93] for birds, iNat21 [91] for animals, fungi, and plants, and FGVC Aircraft [64] for objects. Each sample contains an image and a question prompting the model to identify the specific species or object type(Figure 11). Construction details are as follows:

- We select 100 bird species from CUB-200-2011, and 50 random classes each from the Animal, Fungi, and Plant categories of iNat21, as well as 50 classes from FGVC Aircraft for the Object category. All random selections use a fixed random seed to ensure reproducibility.
- This results in 300 total object ids. For each class, we uniformly sample 30 images.
- For iNat21 entries, we format species names by retaining only the last two words of their taxonomic labels for clarity and consistency.
- For each object class, an 80% training and 20% testing split was established to ensure balanced per-class evaluation.
- For each multiple-choice question, the candidate list includes all 50 or 100 (Object only) class names used in that task split, ensuring consistent, closed-set evaluation.
- The question for each image is: *"What species of bird is in the image? Choose one from below: 1. Cerulean_Warbler, 2. American_Crow, ..., 100. Pine_Warbler"*

OCR [38, 56]. We curate **10.9K** images from three OCR datasets—**COCO-Text** [92], **IIIT5K** [67], and **TextVQA** [26]—each containing word-level bounding box annotations. In each image, a red bounding box highlights the word to be transcribed. The model is prompted to recognize the textual content inside the box based on visual context(Figure 12). The preprocessing steps for dataset creation are summarized below:

- For COCO-Text, we retain only word-level boxes with an area greater than 1500 pixels to ensure sufficient visual resolution.
- For TextVQA, we apply a stricter filtering criterion by retaining only word boxes with area larger than 2000 pixels.
- For IIIT5K, we randomly sample 2,000 images from the original dataset without applying any area-based filtering.
- Each dataset is split into 80% training and 20% validation subsets individually, before merging the resulting splits to form the final OCR benchmark.
- A red bounding box is rendered on each image to highlight the target word location during inference.
- The question for each image: *"What is written in the red bounding box in the image?"*

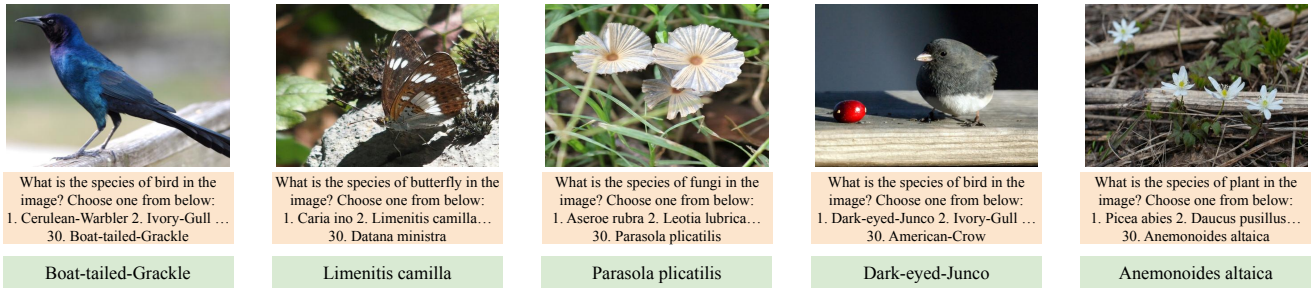


Figure 11. Examples of Fine-grained recognition AVA Samples.

Figure 12. Examples of OCR AVA Samples.

Localization [77, 97]. We curate **34.8K** localization-focused image–question pairs sourced from **Objects365** [78] and **LVIS** [30] (open-domain), **iNaturalist-2021** [91] (birds and animals), and **DIOR** [50] (remote-sensing). Each image contains a single object instance, and the model is prompted to identify its location by providing the bounding box coordinates(Figure 13). The preprocessing steps for dataset creation are summarized below:

- We only retain object instances whose category appears exactly once in an image, to avoid ambiguity in localization supervision.
- We filter out objects with extremely small or large bounding boxes, retaining only those whose normalized area falls within the range of $0.002 < \text{area} < 0.5$ relative to the image.
- For **Objects365** and **LVIS**, we manually select 20 target object categories each. If the number of valid images for a category exceeds 700, we randomly sample 700 using a fixed seed for reproducibility.
- For **iNaturalist-2021**, we select 10 categories from the `aves` (birds) class and 10 from the `mammalia` (mammals) class, following the same filtering and sampling strategy. All scientific names are mapped to common names to improve interpretability and model alignment.
- For **DIOR**, we follow the same filtering steps and manually select 10 target categories, sampling up to 700 images per category as needed.
- All images are padded to square format using a consistent background color computed as the mean RGB value across multiple image processors:

$$\text{background color} = RGB(124, 120, 111)$$

  The padding preserves content aspect ratio and ensures uniform input dimensions across models.
- For each object category, an 80% training and 20% testing split was performed after filtering and sampling, ensuring balanced and fair evaluation.
- The question for each image follows the format: *"Provide bounding box coordinate for red-tailed hawk."* The object name is dynamically replaced depending on the image.

Recognition [22, 89, 112]. We curate **44.9K** recognition samples from four datasets spanning diverse visual domains—**Objects365** [78] and **LVIS** [30] (open-domain objects), **iNaturalist-2021** [91] (birds and animals), and **DIOR** [50] (remote sensing). These samples are derived from the same images and object instances used in the localization benchmark. However, instead of asking for bounding box prediction, the recognition task requires the model to identify the object within a visually highlighted region (Figure 14).

The preprocessing steps for dataset creation are summarized below:

- We apply the same curation strategy as in localization: only one valid instance per image, with normalized bounding box area between 0.2% and 50% of the image.

Atomic Visual Abilities (AVA)	Dataset	Domain	# Train Samples	# Test Samples
Localization	Objects365 [78]	Open	27.9K	6.9K
	LVIS [30]	Open		
	iNaturalist-2021 [91]	Bird, Animal		
	DIOR [50]	Remote-Sensing		
Counting	VQAv2 [27]	Open	10.8K	2.8K
	FSC [74]	Open		
	CARPK [34]	Car		
	Crowd Surveillance [49]	People		
	LVIS [30]	Open		
Fine-grained	CUB-200-2011 [93]	Bird	7.2K	1.8K
	iNaturalist-2021 [91]	Fungi,Plant, Animal		
	FGVC-Aircraft [64]	Object		
Absolute Depth	NYU-Depth V2 [80]	Indoor Scene	6.8K	1.8K
	KITTI [24]	Outdoor Scene		
Relative Depth	NYU-Depth V2 [80]	Indoor Scene	9.2K	2.4K
	KITTI [24]	Outdoor Scene		
OCR	COCO-Text [92]	Open	8.8K	2.2K
	IIIT5K [67]	Open		
	TextVQA [26]	Open		
Orientation	EgoOrientBench [40]	Open	6.9K	1.6K
	CURE-OR [83]	Indoor		
Object Recognition	Objects365 [78]	Open	37.9K	7K
	LVIS [30]	Open		
	iNaturalist-2021 [91]	Bird, Animal		
	DIOR [50]	Remote-Sensing		
Action Recognition	MiT [68]	Open	12K	3K
Texture	DTD [15]	Open	10.6K	2.7K
	Kylberg [46]	Open		
	KTH-TIPS [84]	Open		
	KTH-TIPS2 [65]	Open		
Spatial Reasoning	Objects365 [78]	Open	9.9K	1.6K
	LVIS [30]	Open		
	NYU-Depth V2 [80]	Indoor Scene		
Scene Recognition	Places434 [116]	Open	11.1K	2.8K
	AID [101]	Remote-Sensing		
Emotion	RAF-DB [51]	Human	11.9K	5.1K
	ExpW [53]	Human		
Color	Objects365 [78]	Open	11.2K	2.8K
Total	-	-	182.2K	44.5K

Table 2. **Detailed statistics of** AVA-BENCH**.**

For each dataset, 10 or 20 object categories are manually selected.

- From **Objects365** and **LVIS**, we select 20 object categories each, and randomly sample up to 700 images per category (using a fixed seed for reproducibility).
- From **iNaturalist-2021**, we retain 10 species from the `Aves` (birds) and 10 from `Mammalia` (mammals) branches. Scientific names are converted to common English names for accessibility. Each species contributes up to 700 images.
- From **DIOR**, we select 10 object categories and apply the same filtering and sampling strategy (max 700 images per class).
- All images are padded to square shape using a consistent

Figure 13. Examples of Localization AVA Samples.

background color, computed from the average mean pixel values across nine vision-language processors, to ensure uniform input dimensions.

- For each object category, an 80% training and 20% testing split was performed after filtering and sampling, ensuring balanced and fair evaluation.
- **Unlike localization**, where bounding boxes are not rendered and must be predicted, in recognition the **red bounding box is explicitly drawn** onto each image to guide the model's attention.
- The question format is: *"What is in the red bounding box? Choose from the following option: 1. airport, 2. american robin, ..., 70. vulpes vulpes"* The 70 object categories are shared across datasets and randomly shuffled for each question instance.

Color [12, 94]. We curate **14K** images from **Objects365** [78] and **LVIS** [30], each contributing 7K samples. This AVA focuses on assessing **color perception** in natural scenes. For each image, we extract a coherent color region using the following pipeline:

- We apply SLIC superpixel segmentation and convert the image to LAB color space.
- Superpixels with similar color values are merged to form larger regions of consistent color.
- Among all candidate regions, we select the **top-1 region with the lowest internal color variance** as the final choice.
- A red bounding box is drawn on the selected region, and the most frequent RGB color within this region is used as the answer.
- For each object category, an 80% training and 20% testing split was performed after filtering and sampling, ensuring balanced and fair evaluation.
- The question format for each sample is: "What color is shown within the red bounding box?"

Action [82, 94, 103]. We curate a total of **15K image–question pairs** from the **Moments in Time** [68] dataset, covering a wide range of human actions and activities. Each sample is derived from a short video clip annotated with a specific action label (Figure 15). Construction details are as follows:

- Similar action labels are merged into a unified category for clarity. We have 301 categories in total.
- For each class, we randomly sample up to 40 training videos and 10 testing videos, ensuring broad yet balanced category coverage.
- The middle frame of each selected video is extracted and used as the image representing the associated action.
- Since listing all classes may exceed the token limits of many vision–language models, we randomly sample 100 action options per question, ensuring the ground-truth answer is always included.
- The question format for each sample is: "Which action or activity is shown in the image? Choose from the following option: 1. buying, 2. catching, ..., 100. boxing"

Emotion [48, 104]. We curate **17K image-question pairs** from two large-scale facial expression datasets: **RAF-DB** [51] and **ExpW** [53]. These datasets consist of human portraits labeled with one of seven basic emotions: *happy*, *sad*, *angry*, *fear*, *surprise*, *neutral*, and *disgust*. Each image is annotated with a bounding box localizing the face of interest.

The preprocessing steps for dataset creation are summarized below:

- Emotion labels across datasets were unified by consolidating synonymous terms (e.g., `happiness` and `happy`, `anger` and `angry`) to ensure consistent categorization across all samples.
- Bounding box annotations provided in the datasets were used to highlight the specific individual in multi-person scenes.
- An 80/20 train-test split was applied independently per emotion category to maintain class balance during evaluation.
- The question for each image is framed as: *"Which of the following best describes the person's emotion in the red box? 1. happy, 2. sad, 3. angry, 4. fear, 5. surprise, 6. neutral, 7. disgust."*

Scene [16, 20]. We curate **13.9K image-question pairs** from two diverse datasets: **Places434** [116] (open-domain)

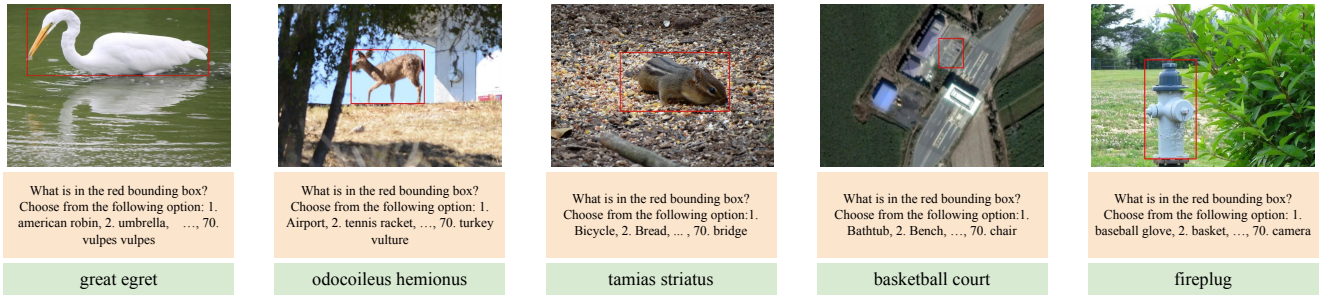


Figure 14. Examples of Recognition AVA Samples.

and **AID** [101] (remote sensing). Each image is paired with a multiple-choice question, where the model selects the correct scene category from a pool of 30 randomly sampled options. The final set includes **463 unique scene classes** spanning a wide range of environments(Figure 16). The preprocessing steps for dataset creation are summarized below:

- GPT-4o was utilized to standardize labels across datasets by converting fine-grained labels into single, unified labels. Humans carefully checked each conversion to merge the newly converted labels conveying the same meaning with existing labels, ensuring semantic clarity and fluency.
- A uniform distribution was maintained by extracting exactly 30 images per scene category, preventing category imbalance and ensuring consistent representation across classes. Classes with less than 30 images were discarded.
- For each scene category within both datasets, an 80% training and 20% testing split was established, ensuring balanced and fair evaluation conditions.
- The question for each pair: "What is the scene class of the image? Choose one from below: 1. Entrance hall, 2. Lawn, ... 30. Snowy Mountain." These 30 options were selected by randomly sampling from the complete set of scene classes within each respective dataset, maintaining diversity and preventing predictable patterns.

Texture [18, 23]. To assess a VFM's ability to distinguish fine-grained visual patterns, we curate **13.2K image-question pairs** from diverse surface textures using close-up images from four open-domain datasets: **DTD** [15] , **Kylberg** [46] , **KTH-TIPS** [84] and **KTH-TIPS2-b** [65]. These datasets encompass a diverse array of texture types—such as *striped*, *aluminum foil*, and *zigzagged*—capturing subtle visual patterns that are essential for accurate texture recognition. Each image is paired with a multiple-choice question, requiring the model to select the correct texture label from a set of options(Figure 17). The preprocessing steps for dataset creation are summarized below:

- Images where textures appeared as part of larger objects in cluttered scenes or within complex real-world photographs were discarded, to ensure that textures were clearly localized and recognizable without contextual interference.
- Each texture attribute was represented by multiple images, with a minimum of 120 and a maximum of 480 samples per attribute, ensuring diversity and preventing memorization of fixed patterns by the models.
- For each texture attribute, an 80% train and 20% test split was ensured, reaching uniform distribution and fair evaluation.
- The question for each pair:*"What is the texture attribute of the image? Choose one from below: 1. banded, 2. blotchy, ..., 47. veined."* The provided options exactly match the entire option pool from each respective dataset and were shuffled to avoid bias.

Orientation [40, 106]. To evaluate viewpoint understanding, we curate **8.5K image-question pairs** from two specialized datasets: **CURE-OR** [83] and **EgoOrient-Bench** [40] . These datasets provide uncluttered images of objects captured from nine distinct orientations—*front*, *back*, *left*, *right*, *top*, *front left*, *front right*, *back left*, and *back right*—allowing models to learn pose-specific cues without requiring bounding boxes(Figure 18). The preprocessing steps for dataset creation are summarized below:

- For EgoOrientBench, each object class was ensured to appear in multiple orientations, with at least 10 and at most 40 samples per orientation label. This encourages models to learn generalized representations rather than memorizing specific arrangements.
- From CURE-OR, we selected object instances photographed against two different background conditions using three distinct capture devices, ensuring variation in imaging style without compromising clarity.
- For each object, 80% of the images from each orientation were assigned to the training set, and the remaining 20% to the test set, ensuring balanced representation and fair evaluation across orientations.
- The question for each pair: *"What is the orientation of the toy plane in the image? Choose one from below: 1. front, 2. front right, 3. right, 4. back right, 5. back, 6.*

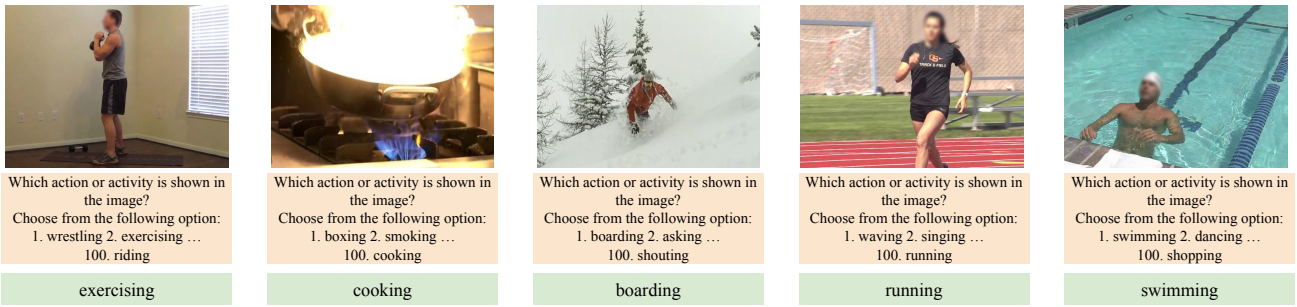


Figure 15. Examples of Action AVA Samples.

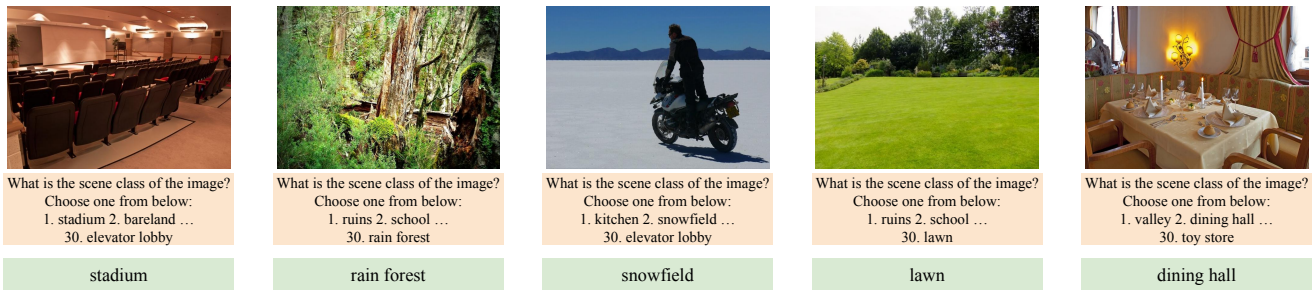


Figure 16. Examples of Scene AVA Samples.

back left, 7. left, 8. front left, 9. top". These nine options represent the common orientation labels provided by both datasets.

Absolute Depth [66, 102, 111]. We curate **9K** image-object pairs from **NYU-Depth V2** [80] for indoor scenes and **KITTI** [24] for outdoor scenes. These datasets contain aligned RGB and depth information. Each image includes an object annotated with a bounding box, and the task requires estimating the absolute depth of the object in meters. This value is then matched against discretized ground-truth bins for evaluation(Figure 19).

The preprocessing steps for **indoor** dataset creation are summarized below:

- Ambiguous object categories were excluded via a manually curated list of label IDs that often lack clear boundaries or meaningful depth interpretations (e.g., wall, floor, ceiling, etc.).
- Images were resized to a fixed resolution of 384×384, padding vertically as needed to preserve the original aspect ratio.
- A minimum bounding box area threshold was enforced after resizing all images. Specifically, we filtered out bounding boxes smaller than 500 pixels to ensure sufficient spatial resolution for the model.
- To ensure depth variation and avoid trivial samples, only object classes with at least 3 distinct depth bins (i.e., meaningful distribution over depth) were retained.
- For each label, depth bins were required to have a minimum of 10 and a maximum of 30 image samples. We removed bins with insufficient data to meet this requirement and capped those with excess samples by sorting instances based on bounding box area.
- After filtering, we retained 45 object classes, resulting in 4.4K unique image-object pairs. The dataset was split into 80% train and 20% test, preserving label and bin balance.
- The question for each pair: *"From the camera's perspective, estimate how far the closest point of the cabinet (highlighted by a red box) is from the camera in real-world distance, in meters. Select the best answer from the options below: A. 1-2, B. 2-3, C. 3-4, D. 4-5, E. 5-6, F. 6-7"*.

The preprocessing steps for **outdoor** dataset creation are summarized below:

- To ensure the depth estimation task remains non-trivial, only objects with a minimum distance of 8 meters from the camera were considered. This avoids bias toward near-field predictions and better evaluates model precision in far-range perception.
- Depth values were discretized into bins, and for each class, we selected between 20 and 60 samples per bin to ensure coverage while avoiding overrepresentation. Bins with fewer than 20 samples were discarded. When bins exceeded 60 samples, selection was sorted by bounding box area to prioritize larger, more reliable objects.
- Image crops were extracted per object while maintain-

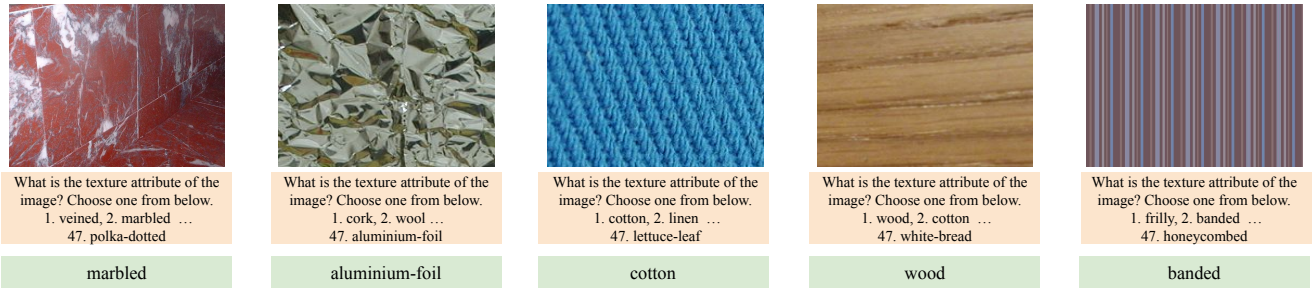


Figure 17. Examples of Texture AVA Samples.

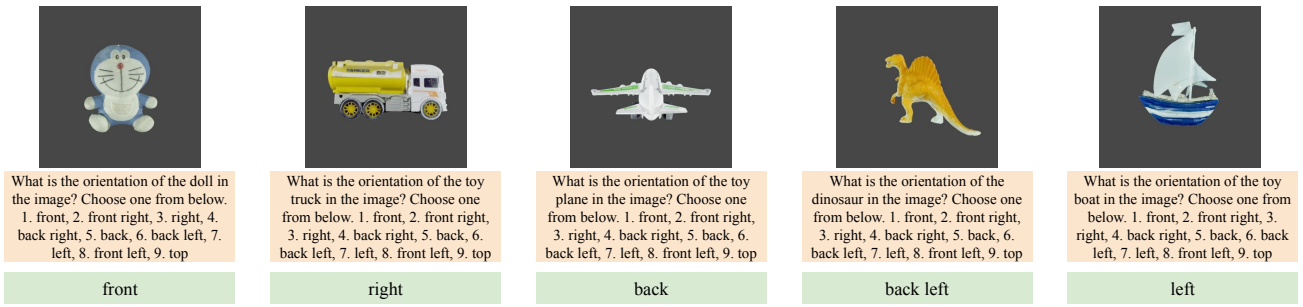


Figure 18. Examples of Orientation AVA Samples.

ing the following aspect ratio constraint to preserve visual consistency: the width of the crop must be within the range $[\text{height}, 2 \times \text{height}]$. This was enforced using the original image dimensions before padding or resizing.

- Objects touching any edge of the image were excluded to reduce the likelihood of partial occlusion or clipping.
- Images were padded vertically as needed to preserve the original aspect ratio.
- The final outdoor absolute depth set contains approximately 6K samples. For each object class and depth bin, an 80/20 train-test split was applied to maintain consistency in evaluation.
- The question for each outdoor sample: *"Estimate the distance from the camera to the closest part of the cyclist (highlighted by a red box) in meters. Choose the best option: A. 8-9, B. 10-11,..., H. 30-31."*

Relative Depth [66, 102, 111]. We curate **11.6K** image-object pairs from **NYU-Depth V2** [80] for indoor scenes and **KITTI** [24] for outdoor scenes, targeting the task of identifying which of two objects in an image is closer to the camera. Each image contains two distinct objects, each annotated with a bounding box. The model is asked to compare their absolute depth and choose the object that appears closer to the camera (Figure 20).

The preprocessing steps for **indoor** dataset creation are summarized below:

- Ambiguous object categories were excluded via a manually curated list of label IDs that often lack clear boundaries or meaningful depth interpretations (e.g., wall, floor, ceiling, etc.).
- Only object pairs with valid bounding boxes (i.e., non-overlapping, fully inside image boundaries) were considered.
- To ensure perceptual clarity, candidate object pairs were filtered by requiring an absolute depth difference of at least **0.5 meters** between them.
- For a given object class, only those with at least **10** valid pairings were retained, and a maximum of **30** total pairings per label were allowed.
- After filtering and sampling, we retained **131** object pairs across **5.7K** total questions. Images were padded vertically as needed to preserve their original aspect ratio. These were split into 80% train and 20% test sets while maintaining object label and depth-difference balance.
- Each question is posed as: *"Estimate the real-world distances between the objects in this image. Which object is closer the camera, the sink (highlighted by a red box) or the towel (highlighted by a blue box) to the camera? Choose one option from below: 1. red, 2. blue"*.

The preprocessing steps for **outdoor** dataset creation are summarized below:

- Object annotations were sourced for the following categories: `Car`, `Van`, `Pedestrian` (merged with `Person sitting`), and `Cyclist`.
- Pairs were formed using both intra-class (e.g., Car vs. Car) and inter-class (e.g., Car vs. Pedestrian) combina-

From the camera's perspective, estimate how far the closest point of the car (highlighted by a red box) is from the camera in real-world distance, in meters. Select the best answer from the options below: A. 8-9, B. 9-10, … , AM.46-47

9-10

From the camera's perspective, estimate how far the closest point of the monitor (highlighted by a red box) is from the camera in real-world distance, in meters. Select the best answer from the options below: A. 1-2, B. 2-3, C. 3-4, D. 4-5, E. 5-6

1-2

From the camera's perspective, estimate how far the closest point of the pedestrian (highlighted by a red box) is from the camera in real-world distance, in meters. Select the best answer from the options below: A. 8-9, B. 9-10, ... , N. 21-22

12-13

Figure 19. Examples of Absolute Depth AVA Samples.

tions.

- Pairs were retained only if the depth difference between the two objects was at least 0.5 meters, ensuring a meaningful perceptual gap.
- To avoid ambiguity and incomplete visual evidence, the following filters were applied:
  - Pairs with occluded objects were excluded.
  - Pairs where either object was touching the image edge were discarded.
  - Only crops satisfying the aspect ratio constraint height $\leq$ width $\leq 2 \times$ height were included, ensuring visual consistency.
  - Images were padded vertically as needed to preserve their original aspect ratio.
- After filtering, approximately 6K valid image-object pairs were retained. For each pair type, an **80/20 train-test split** was applied while maintaining distributional balance over object categories and depth separations.
- Each question is posed as: *"Which object is closer to the camera, the van (highlighted in red) or the cyclist (highlighted in blue)? Choose one: 1. red, 2. blue."*

B. Experiment Details

B.1. Hyperparameter Details

To ensure reproducibility and fairness, we carefully followed the official TinyLLaVA hyperparameter recommendations for stages 1 and 2, maintaining both the global batch size and learning rate as prescribed (in Table 4.). For stage 3, which incorporates LoRA-based fine-tuning, we selected a learning rate of 1e-4 and explored multiple LoRA dimensions (64, 128, 256). To validate these choices, we conducted preliminary experiments using three representative VFMs: DINOv2, CLIP, and SigLIP-2. We evaluated performance on two representative AVA tasks (OCR and Recognition), as summarized in Table 3.

The results consistently show that the recommended learning rate of 1e-4 yields stable and strong performance, whereas alternative learning rates often underperform or lead to instability. Similarly, LoRA dimensions between 64 and 128 produce comparable and reliable results, while extreme values show diminishing returns. Based on these observations, we adopt the recommended configuration (learning rate 1e-4, LoRA dimension 128) throughout our experiments. The overall hyperparameters of Stage-1 vision-language alignment pretraining, Stage-2 visual instruction tuning and Stage-3 AVA-BENCH evaluation are shown in Table 4.

B.2. Metric Details

Color Recognition. We use CIEDE2000 [58] to calculate the color differences using the `colour` Python library. Specifically, we convert the predictions and ground-truths from CIE XYZ tristimulus format to CIE L*a*b* colour space with `colour.XYZ_to_Lab`, followed by the `colour.delta_E(pred, gt, method="CIE 2000")` for color differences.

Absolute Depth & Counting. We use the mean absolute error relative to the ground-truth (see Equation 1). This normalization ensures that errors involving greater distances or counts, which are inherently more challenging, are proportionally penalized less severely.

$$\text{MAE/GT} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\, y_i - \hat{y}_i \,\right|}{y_i}. \tag{1}$$

where N is the number of test samples, y_i is the ground-truth (depth or count) for sample i and $\hat{y}_i$ the model prediction.

Localization. We use the Generalized Intersection-over-Union (GIoU [76], Equation 2):

$$\text{GIoU}(A,B) = \frac{|A \cap B|}{|A \cup B|} - \frac{\left|\, C \setminus (A \cup B) \,\right|}{|C|}. \tag{2}$$

where A and B are the prediction and ground-truth bounding boxes and C is the smallest (axis–aligned) enclosing box of $A \cup B$.

Task	Model	lr 1e-5	lr 1e-4	lr 5e-4	LoRA 64	LoRA 128	LoRA 256
OCR	DINOv2	7.26	9.97	10.6	10.85	9.97	10.98
	SigLIP-2	79.68	81.18	77.63	81.51	81.18	81.25
	CLIP	54.23	60.44	60.9	60.79	60.44	61.73
Recognition	DINOv2	83.92	86.31	unstable	86.39	86.31	86.46
	SigLIP-2	87.04	88.19	unstable	88.42	88.19	88.36
	CLIP	83.08	85.02	unstable	84.79	85.02	84.18

Table 3. Hyperparameter exploration for OCR and Recognition tasks using three representative VFMs. Results are reported across different learning rates and LoRA dimensions.

	TinyLLaVa		**AVA-BENCH**
Hyperparameter	Stage 1	Stage 2	Stage 3
batch size	16	4	4
grad accum steps	4	8	1
LR	1e-3	2e-5	1e-4
LR schedule	cosine decay		cosine decay
LR warmup ratio	0.03		0.03
weight decay	0		0
epoch	1		10 (20 for localization AVA)
optimizer	AdamW		AdamW
DeepSpeed stage	3		3
components finetuned	Connector	Connector + LLM	Connector + LoRA on LLM
sample size	558K	665K	–

Table 4. Hyperparameters of TinyLLaVa and AVA-BENCH Evaluation Stage

OCR. We evaluate OCR performance with Average Normalized Levenshtein Similarity (ANLS) [4]:

$$\begin{aligned} \mathrm{NLS}(p, g) &= 1 - \frac{\mathrm{Lev}(p, g)}{\max(|p|, |g|)}, \\ \mathrm{ANLS} &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{NLS}(p_i, g_i) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\mathrm{Lev}(p_i, g_i)}{\max(|p_i|, |g_i|)} \right). \end{aligned} \tag{3}$$

where $\mathrm{Lev}(p, g)$ is the (Levenshtein) edit distance and $|.|$ denotes string length and N is the number of testing samples.

Others. All other AVAs employ standard accuracy metrics.

C. More Results and Analysis

C.1. Detailed Overall Results

The results used for plotting Figure 7 and Figure 8 are presented in Table 5 where the best performance for each is AVA is **bold** and the second best is in *italics*. For each VFM, the first row is the performance, and the second row is the rank.

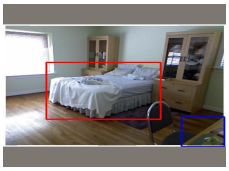
C.2. Detailed Analyses for Each AVA

In the main results, we reported the overall performance of VFMs across various AVAs. However, aggregate metrics can sometimes obscure important nuanced insights. To gain deeper understanding, we conduct detailed analyses by partitioning test samples based on specific criteria (e.g., object size in localization tasks) and examining whether these subgroup trends align with overall performance. Generally, the detailed analyses affirm overall trends, but notable exceptions exist, particularly in localization.

Localization. We split localization testing samples based on normalized bounding box sizes (relative to image size), where 0.1 indicates an object occupies 10% of the image area. As illustrated in Figure 8 (b), VFMs surprisingly exhibit minimal performance differences when localizing large objects (0.3–0.5). Conversely, performance disparities amplify as object size decreases, revealing significant weaknesses in MiDas and SAM for smaller objects. Consequently, the lower overall performance of MiDas and SAM is predominantly due to poor handling of small targets. Practitioners should thus consider object size distributions

Estimate the real-world distances between the objects in this image. Which object is closer the camera, the computer (highlighted by a red box) or the paper (highlighted by a blue box) to the camera? Choose one option from below: 1. red, 2. blue

blue



Estimate the real-world distances between the objects in this image. Which object is closer the camera, the bed (highlighted by a red box) or the desk (highlighted by a blue box) to the camera? Choose one option from below: 1. red, 2. blue

blue

Estimate the real-world distances between the objects in this image. Which object is closer to the camera, the car (highlighted by a red box) or the car (highlighted by a blue box) to the camera? choose one option from below: 1. red , 2. blue

red

Figure 20. Examples of Relative Depth AVA Samples.

AVA	Abs. Depth	Rel. Depth	OCR	Counting	Localization	Object	Fine-grained	Scene	Action	Spatial	Emotion	Orientation	Texture	Color	Average
Metric	MAE/GT↓	ACC↑	ANLS↑	MAE/GT↓	GIOU↑	ACC↑	ACC↑	ACC↑	ACC↑	ACC↑	ACC↑	ACC↑	ACC↑	CIEDE2000↓	Ranking
SigLIP-2	0.0843	*97.38*	**81.18**	**0.225**	**0.6738**	**88.19**	90.17	72.60	*45.33*	*99.50*	58.39	79.2	*94.08*	*12.25*	2.4
	3	2	1	1	1	1	3	3	2	2	5	4	2	2	
AIMv2	0.1008	95.71	62.44	0.254	0.5896	85.04	**91.78**	*72.86*	43.61	99.13	**60.68**	79.71	**94.11**	19.61	3.9
	9	8	3	3	5	4	1	2	3	6	1	3	1	6	
SigLIP-1	0.0953	96.58	*80.3*	*0.229*	0.6103	*87.84*	*90.94*	**73.4**	**45.76**	99.07	59.6	78.91	93.66	12.64	3.6
	6	7	2	2	4	2	2	1	1	7	4	5	3	3	
CLIP	0.08461	97.04	60.44	0.290	0.5787	85.02	86.83	72.32	41.59	99.25	*60.42*	77.77	93.25	19.85	4.9
	4	4	6	7	7	5	4	5	4	5	2	6	5	7	
InternVL-2.5	*0.08212*	97.00	60.88	0.269	0.5850	83.19	72.83	71.63	36.13	99.5	59.99	75.71	91.21	20.09	5.4
	2	5	5	5	6	7	7	6	7	4	3	7	7	8	
RADIOv2.1	**0.07645**	**97.92**	62.44	0.257	*0.6617*	84.90	85.44	72.50	38.77	**99.69**	56.59	*83.71*	93.32	17.16	3.6
	1	1	4	4	2	6	6	4	5	1	6	2	4	5	
DINOv2	0.08469	97.25	9.97	0.272	0.6598	86.31	85.5	70.99	37.45	99.50	54.44	**85.54**	93.06	21.54	5.3
	5	3	7	6	3	3	5	7	6	3	7	1	6	9	
SAM	0.09792	94.13	9.79	0.313	0.5216	76.68	40.06	58.28	17.25	90.22	36.88	69.37	81.02	**9.87**	7.8
	8	9	8	8	8	8	8	9	9	8	9	8	9	1	
MiDas-3.0	0.09563	96.63	7.72	0.336	0.4490	75.05	32.83	60.19	18.25	53.58	40.40	67.37	86.38	13.28	8
	7	6	9	9	9	9	9	8	8	9	8	9	8	4	

Table 5. The detailed overall results where the best performance for each is AVA is **bold** and the second best is in *italics*. For each VFM, the first row is the performance, and the second row is the rank. Arrows indicate whether lower (↓) or higher (↑) is better.

when selecting VFMs; SAM and MiDas remain viable if target objects are predominantly large.

Counting. Counting performance is generally consistent across different datasets and count ranges (Figure 21). A notable exception is SAM, whose accuracy notably improves in denser scenarios.

Emotion. Emotion recognition results exhibit remarkable consistency, with rankings and relative performances highly stable across emotion categories (see Figure 22).

Orientation. Orientation performance remains consistent overall, with some intriguing exceptions. Specifically, VFMs universally achieve near-perfect accuracy for top-view images, presumably due to the distinctive nature of this viewpoint compared to side or frontal views (see Figure 23).

Absolute Depth. Overall, absolute depth performance is stable, though specific VFMs exhibit distinctive trends. SAM notably struggles with near objects but improves significantly with increased distance. Conversely, RADIO demonstrates an opposite pattern, excelling with nearer objects but deteriorating with greater distances (see Figure 24).

Relative Depth. Relative depth estimation shows uniformly high performance across VFMs, consistently surpassing 90% accuracy. SAM, however, underperforms notably in interior scenes, consistent with the earlier observation in absolute depth and counting that SAM handles smaller, exterior objects better (see Figure 25).

Fine-grained Classification. Fine-grained classification results are consistently robust across datasets, with the exception of SAM and MiDas, both of which are known to lack semantically rich features [9, 19], resulting in poorer performance (see Figure 26).

Scene Recognition. Scene recognition performance is uniformly consistent across all evaluated datasets, echoing the patterns observed in fine-grained classification, where SAM and MiDas again lag behind other VFMs (see Figure 27).

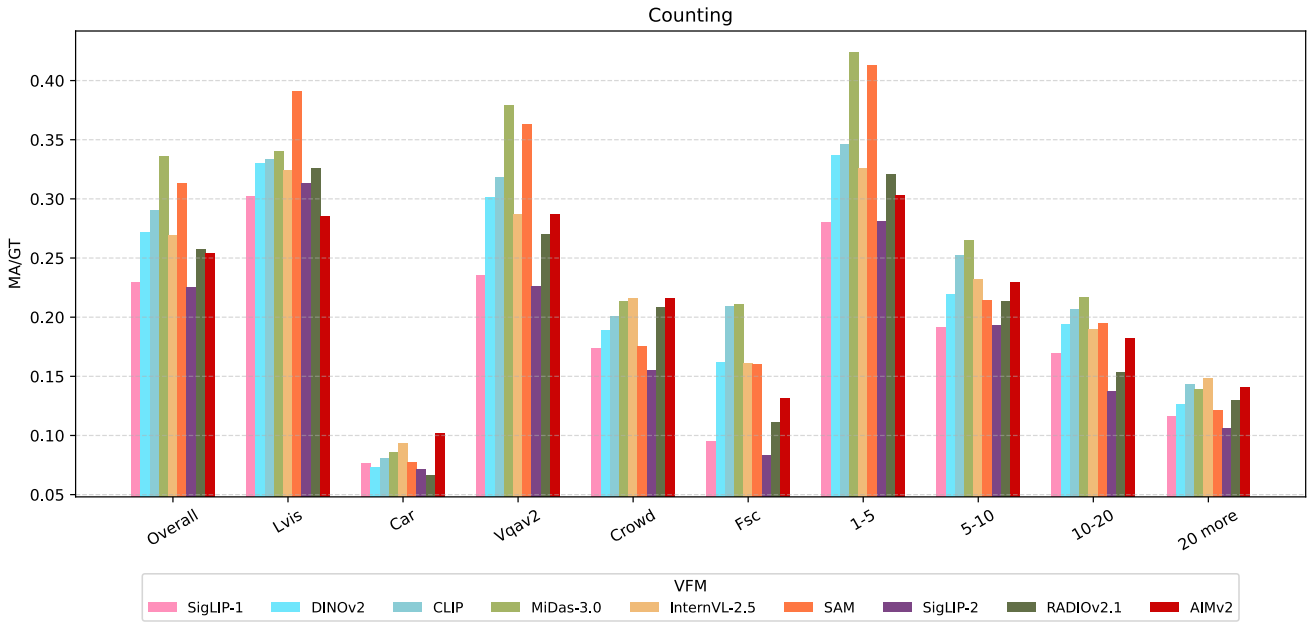


Figure 21. Detail results for counting for overall and different splits based on datasets and ground-truth count range. Lower MAE/GT is better.

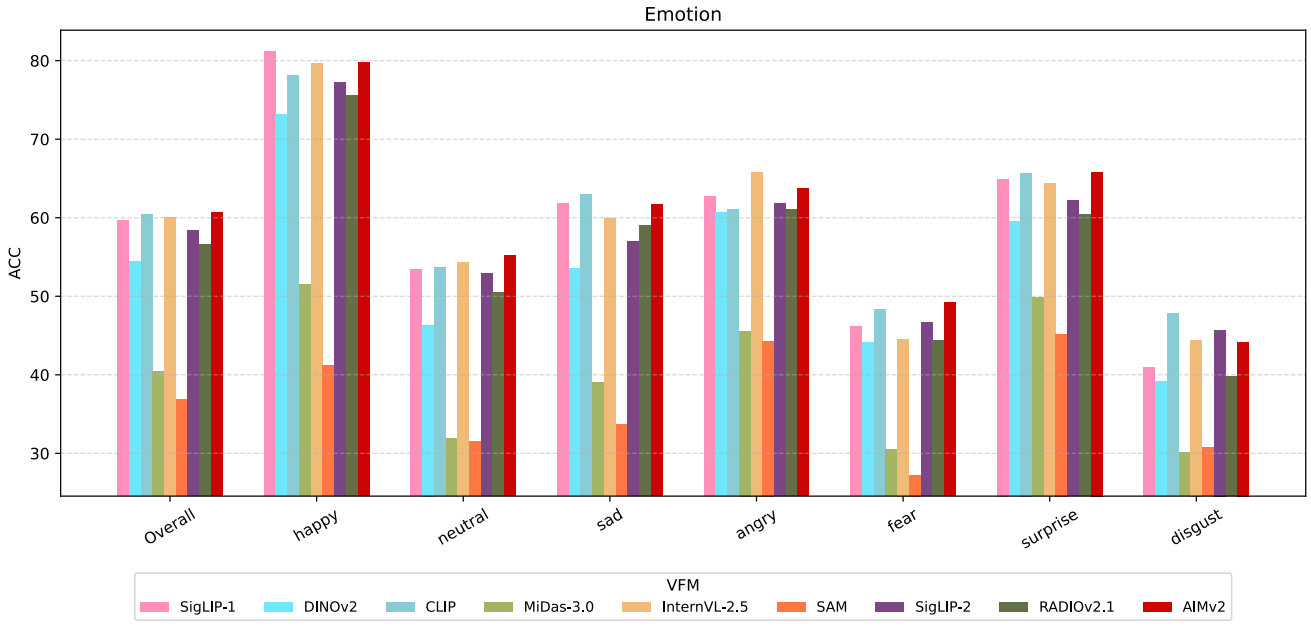


Figure 22. Detail results for emotion for overall and different splits based on emotion types.

OCR. OCR results show consistent patterns with those reported in Section 5.2, highlighting significant underperformance by non-language-aligned VFMs, such as DINOv2 and SAM. Notably, we observe that relative performances across VFMs are stable on short texts (length < 20). However, performance for CLIP and AIM sharply declines with longer text sequences (length > 20), indicating potential limitations in handling extensive textual information (see Figure 28).

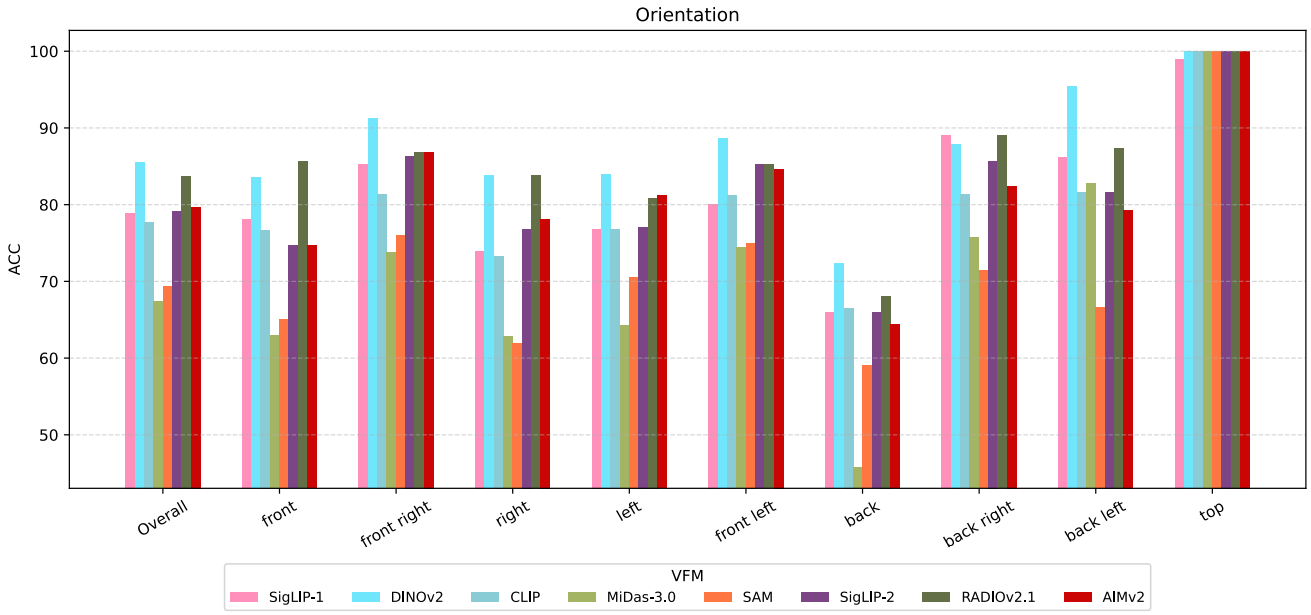


Figure 23. Detail results for orientation for overall and different splits based on viewpoint directions.

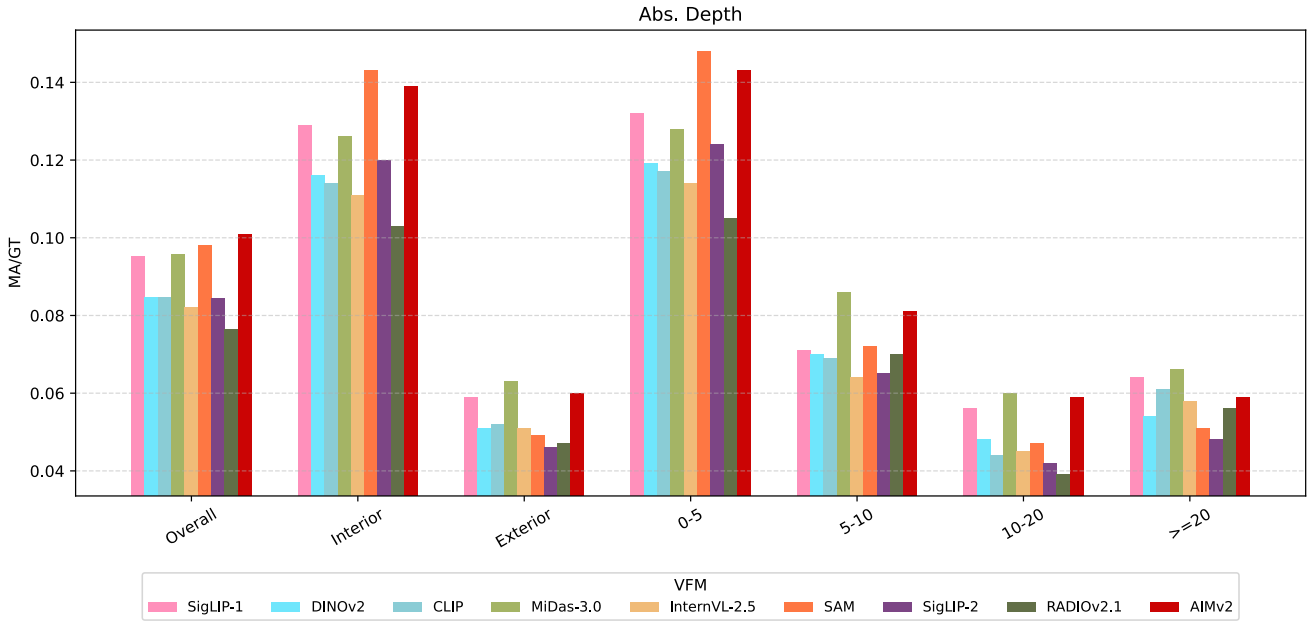


Figure 24. Detail results for absolute depth for overall and different splits based on scene type and object count range. Lower MAE/GT is better.

D. VFM Details

Table 6 provides a detailed overview of the vision foundation models (VFMs) evaluated in our study. For each model, we list its architecture, parameter count, and training data, and we further summarize the training methodology in terms of supervision type, process, and loss functions.

E. Related Works

E.1. VFM Evaluation

Existing evaluation VFM protocols generally fall into two categories. The first focuses on task-specific capabilities, typically attaching tailored heads to VFMs, followed by

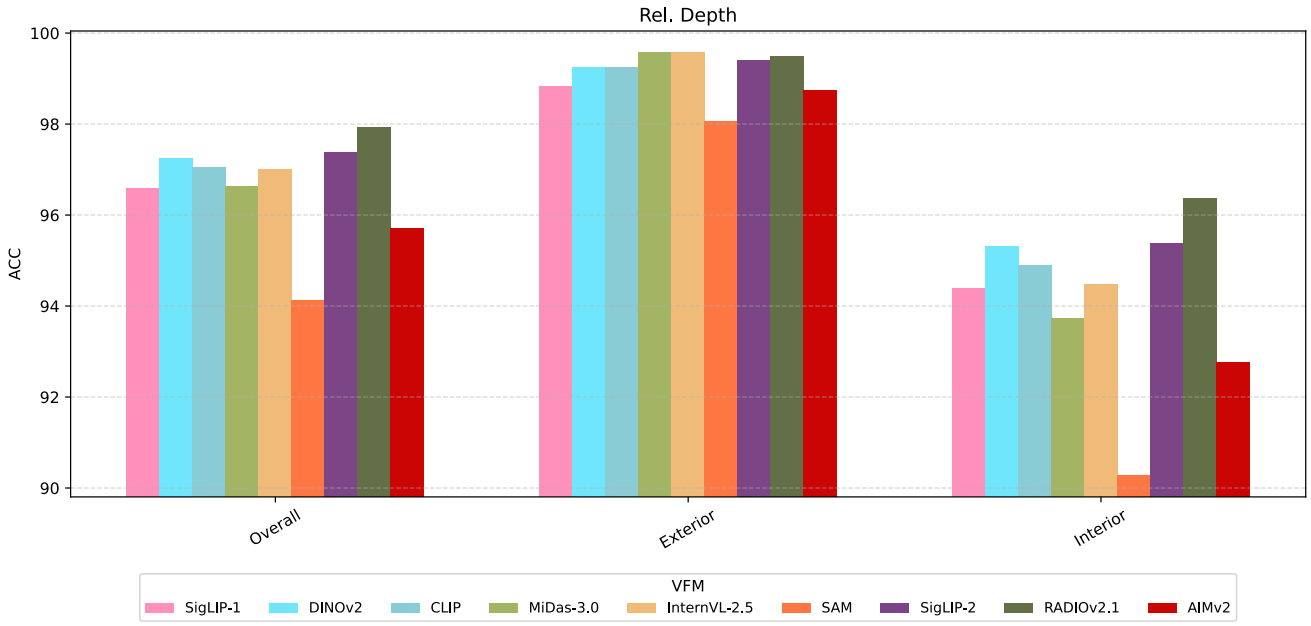


Figure 25. Detail results for relative depth for overall and different splits based on scene type.

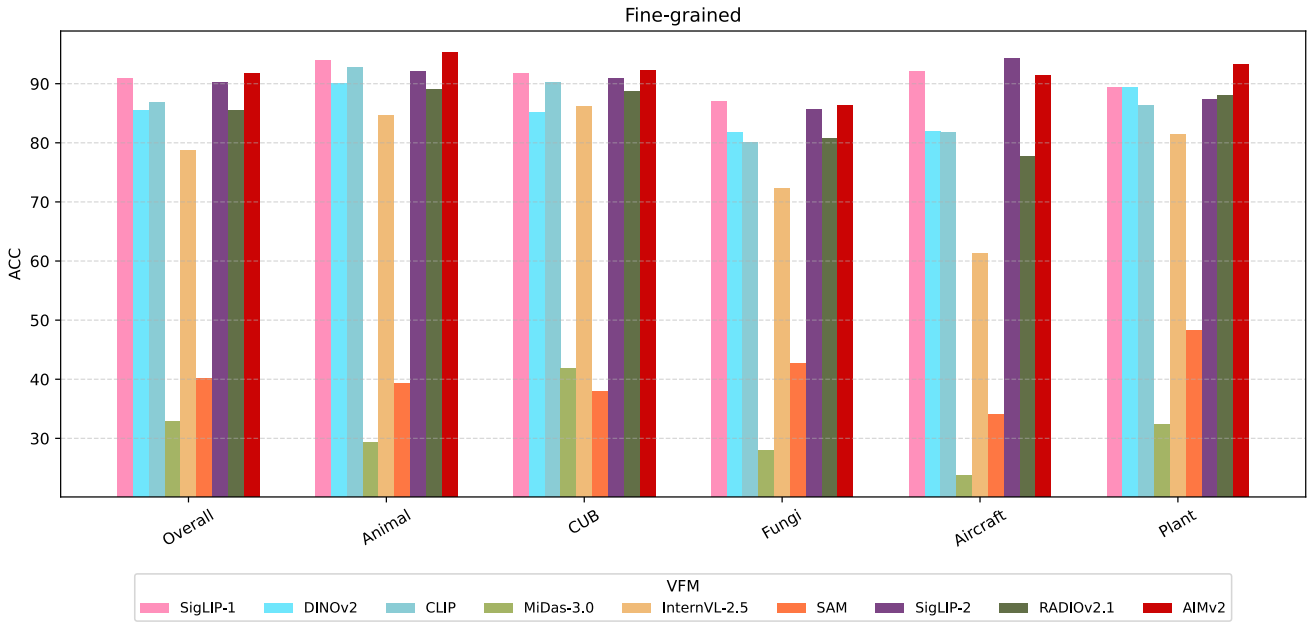


Figure 26. Detail results for fine-grained for overall and different splits based on dataset type.

fine-tuning and evaluation on dedicated datasets such as ImageNet for classification [31] and COCO for detection or segmentation [85]. For example, DINOv2 [70] is evaluated on image and video classification, instance recognition, image retrieval, semantic segmentation, and depth estimation. For each task, a task-specific head is trained.

To better capture the diverse and complex perception challenges of the real world, recent studies advocate a more generic approach that leverages large language models (LLMs) as general-purpose heads, evaluating VFMs on broad Visual Question Answering (VQA) benchmarks [13, 55, 117]. For example, in addition to the traditional task-specific evaluation, AIMv2 [21] and RADIO [75] follow the LLM-based evaluation and use a Llama-3.0(8B) [28]

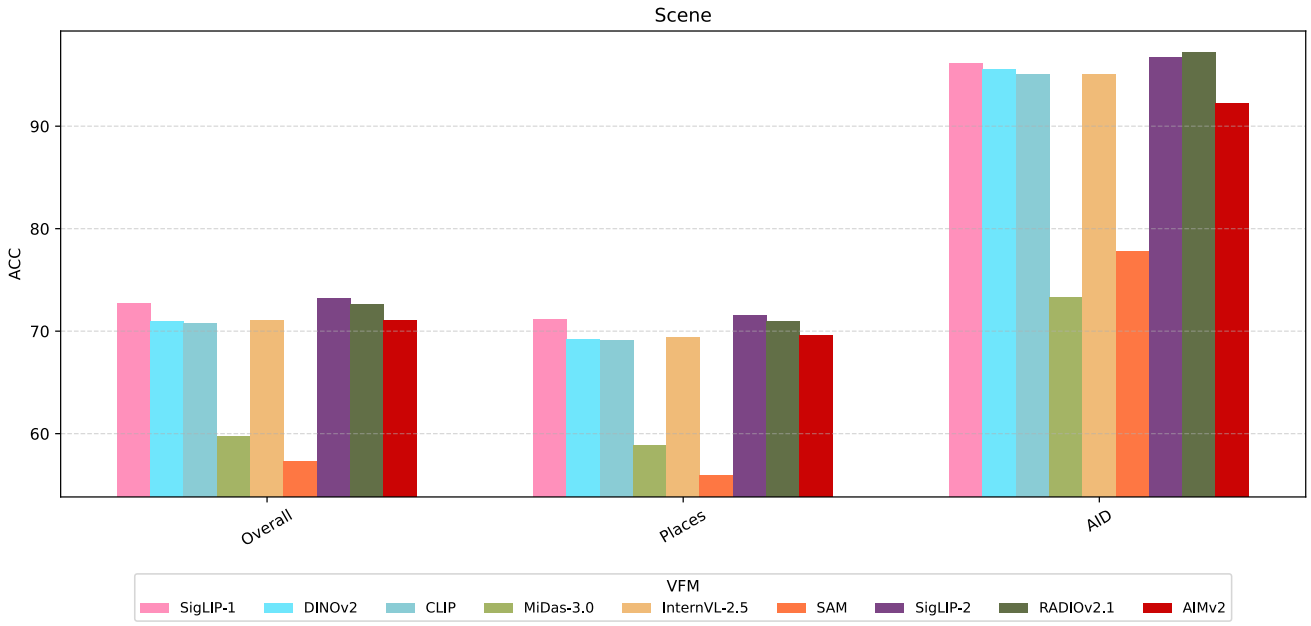


Figure 27. Detail results for scene for overall and different splits based on datasets.

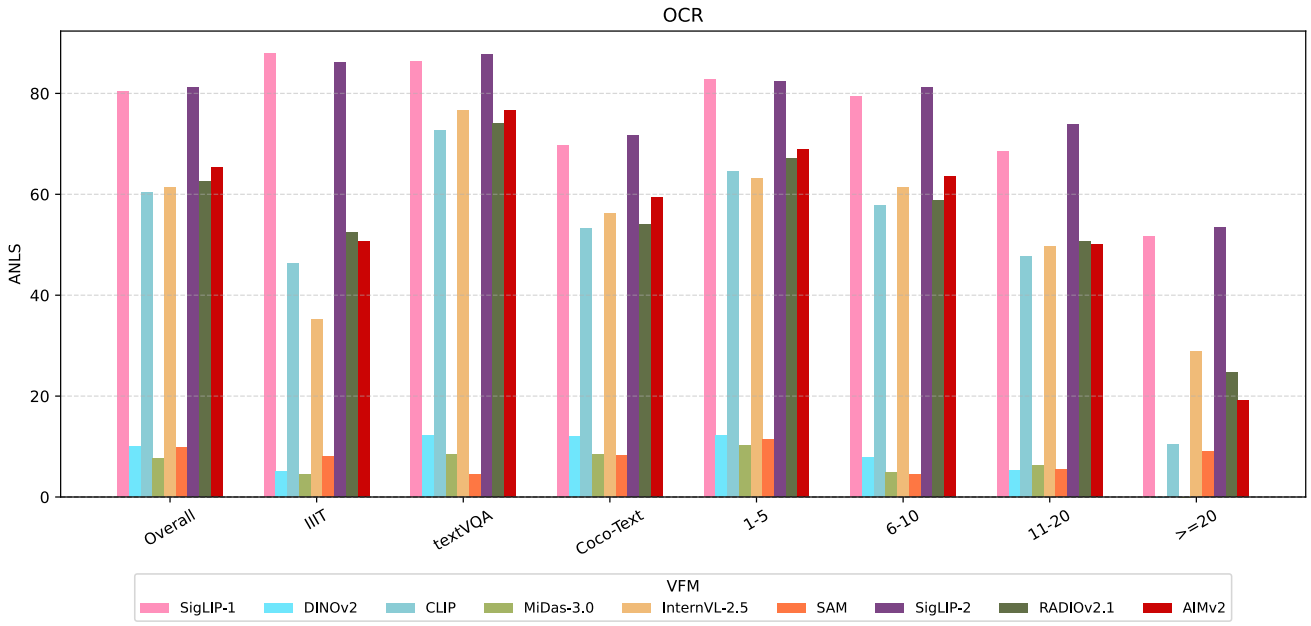


Figure 28. Detail results for counting for overall and different splits based on dataset domain and character length. Higher ANLS is better.

and a Vicuna-1.5(7B) [115] respectively, demonstrating a shift towards generalized multimodal evaluation.

E.2. Atomic Visual Abilities

As discussed in subsection A.1, foundational visual concepts—such as number, color, texture, object identity, and spatial relations—have long been recognized as crucial building blocks in compositional Text-to-Image (T2I) benchmarks. Given their foundational role in generation tasks, these primitives naturally underpin perceptual tasks as well. For example, a concept like 'number' directly translates into the perceptual task of counting.

A recent work [7] introduced AVSBench to evaluate whether MLLMs understand basic *geometric* features,

VFM	Architecture	Parameters	Training Data	Training Details
SigLIP-2	Dual-tower ViT encoders for image and text embeddings with MAP pooling layers	So400m (400M)	WebLI dataset	•Supervision: Supervised (image-text pairs) •Process: Pretrained on 40B samples, large-batch training (32k) •Loss: Pairwise sigmoid ITC + captioning/grounding
AIMv2	ViT-based vision encoder with an autoregressive multimodal decoder for patch and token reconstruction	AIMv2-Huge (600M)	DFN, COYO, HQITP	•Supervision: Supervised (multimodal autoregressive) •Process: Pretrain (224 px) → finetune (336/448 px), long training •Loss: Joint reconstruction of image patches and text tokens
SigLIP-1	Dual-encoder with independent ViT and text transformer projecting to a shared embedding space	So400m (400M)	WebLI dataset	•Supervision: Supervised (image-text pairs) •Process: Large-scale pretraining, efficient setup with 32k batch •Loss: Pairwise sigmoid contrastive loss
CLIP	Dual-tower model using a ViT image encoder and Transformer text encoder for contrastive alignment	ViT-L/14 (428M)	Internet-collected dataset	•Supervision: Supervised (image-text pairs) •Process: Weeks-long pretraining on multi-GPU/TPU with large batches (≥32k) •Loss: Contrastive InfoNCE
InternVL-2.5	Large multimodal architecture combining a high-capacity ViT encoder with an LLM for image-text reasoning	InternVL2.5 (304M)	FaceCaption, GQA, ChartQA, Many other datasets	•Supervision: Supervised (multimodal LLM) •Process: Pretrained and finetuned on diverse datasets •Loss: Autoregressive next-token + alignment losses
RADIO v2.1	ViT backbone with conditional positional encoding and multi-teacher feature distillation layers	RADIO-Huge (653M)	DataComp1B dataset	•Supervision: Supervised (teacher-student distillation) •Process: 600k steps, AdamW (WD=1e-4), batch scaling law (eff. BS 1024) •Loss: Multi-teacher distillation from CLIP, DINOv2, SAM-H
DINOv2	ViT backbone with patch embedding and projection heads for self-supervised feature representation learning	ViT-Large (300M)	LVD-142M dataset	•Supervision: Self-supervised (no labels) •Process: Teacher-student distillation pipeline with deduplication and retrieval •Loss: Self-distillation contrastive objective
SAM	MAE-pretrained ViT-H image encoder paired with a prompt encoder handling points, boxes, masks, and text queries	ViT-H (637M)	SA-1B dataset	•Supervision: Supervised (segmentation masks) •Process: Pretrained encoder with MAE, promptable training on SA-1B •Loss: Segmentation mask prediction loss
MiDaS-3.0	Multi-scale ResNet-based encoder-decoder network designed for monocular depth prediction from single images	ResNet-Encoder (123M)	DIML Indoor, MegaDepth, ReDWeb, WSVD	•Supervision: Supervised (depth ground truth) •Process: Multi-dataset pretraining, 60 epochs, Adam optimizer with different LR for new vs pretrained layers •Loss: Trimmed MAE (20%) + gradient regularizer

Table 6. Details of Vision foundation model (VFM) used, including architecture, parameter scale, training data, and training procedures.

including angle, boundary, orthogonality, and curvature, which they refer to as atomic visual skills. However, AVS-Bench primarily targets geometric comprehension abilities required for geometric diagrams arising in high-school level mathematics. Moreover, AVSBench provides only test data for MLLMs without addressing potential mismatches between training and test data distributions—an issue highlighted in Section 2.2. Consequently, mispredictions in AVSBench evaluations may result from data distribution mismatches rather than genuine visual deficiencies in VFM. In contrast, AVA-BENCH explicitly emphasizes atomic visual abilities essential for general visual reasoning tasks commonly encountered in real-world scenarios. By aligning training and evaluation data distributions, AVA-BENCH ensures that evaluation outcomes reliably reflect genuine visual perceptual capabilities of VFMs.

Additionally, a concurrent work [99] defines a set of atomic visual capabilities analogous to ours. However, their goal fundamentally differs from ours: [99] aims to build a visual compositional tuning data recipe that builds complex capabilities from simple atomic capabilities, which can significantly reduce instruction-tuning data volume while maintaining strong performance. In contrast, AVA-BENCH's objective is to systematically evaluate VFMs against atomic visual abilities, pinpointing their exact strengths and weaknesses, and providing a comprehensive diagnostic tool to advance the continual development [57, 59–61, 79] of robust vision foundation models.

F. Evaluation efficiency

An important advantage of our framework is its efficiency compared to prior LLM-based evaluation protocols. As summarized in Table 7, existing methods typically rely on large language models such as Vicuna-7B and require ≈230 A100 GPU hours per vision foundation model (VFM). By contrast, our approach adopts a lightweight 0.5B LLM and smaller training data (1.2M samples in total for stages 1 and 2), which reduces the cost to ≈28 A100 GPU hours while still preserving consistent and reliable VFM rankings. This design choice enables practical scaling to a wide range of models without incurring prohibitive resource demands.

Protocol	LLM size	Stage 1&2 data	Stage 1&2 cost	Stage 3
Baseline [1]	Vicuna-7B	1.9M	≈230 A100 h	n/a
AVA-BENCH	Qwen2-0.5B+LoRa(stage 3)	1.2M	≈28 A100 h	Each AVA: avg 4 A100 h

Table 7. Evaluation Efficiency Table

Dataset	Copyright	License
Object365	Objects365 Consortium	CC By 4.0
LVIS	LVIS Consortium	CC By 4.0
iNaturalist-2021	iNaturalist (Terms of Service)	MIT
DIOR	N/A	N/A
VQAv2	VQA Consortium	CC BY 4.0
FSC	CVLab at StonyBrook	MIT
CARPK	Original image owners (PUCPR/PKLot)	N/A
Crowd Surveillance Dataset	N/A	N/A
CUB-200-2011	Annotations: Catherine Wah et al.; images: original owners	CC0 (Public Domain)
FGVC-Aircraft	Annotations: S. Maji et al.; images: original owners	Research only (Non-commercial)
KITTI	Andreas Geiger, Philip Lenz, Christoph Stiller, Raquel Urtasun	CC BY-NC-SA 3.0
NYU-DepthV2	N/A	N/A
coco-text	SE(3) Computer Vision Group, Cornell Tech	CC BY 4.0
IIIT5K	IIIT Hyderabad (annotations); images: original owners	N/A
TextVQA	VQA Consortium	CC BY 4.0
EgoOrientBench	N/A	N/A
CURE-OR	OLIVES at Georgia Institute of Technology	MIT
Moment_int_time	Moments in Time authors	Research/Educational only
DTD	N/A	N/A
KTH-TIPS	N/A	N/A
KTH-TIPS2	N/A	N/A
Places365	MIT CSAIL, Bolei Zhou; images: original owners	MIT (code); images original copyright owners
AID	AID authors (Gui-Song Xia et al.); images from Google Earth providers	N/A
RAF-DB	N/A	N/A
ExpW	N/A	N/A

Table 8. Dataset copyright and licensing information for all datasets used in AVA-BENCH

For stage 3, our framework further leverages LoRA-based fine-tuning, where each AVA is trained on only 6K–10K samples. This procedure requires ≈4 A100 GPU hours per AVA on average, making it highly lightweight compared to full model finetuning. In summary, AVA-BENCH achieves a more diagnostic evaluation with considerably lower overhead than prior work.

G. Dataset Copyright/License

To ensure ethical and legal use of datasets, we summarize the copyright and licensing information of all benchmarks employed in our experiments (Table 8). The majority of the datasets we use are publicly available under open licenses such as CC BY 4.0, MIT, or CC0, which permit research and redistribution with proper attribution. Some datasets (e.g., FGVC-Aircraft, Moments in Time) are restricted to research-only or educational use, and we adhered to these conditions. For datasets without explicit licensing details, we used them strictly within the scope of non-commercial academic research.