Concept-skill-based Curriculum Learning for Large Vison-Language Models

Jaewoo Lee
jwlee8877@gmail.com

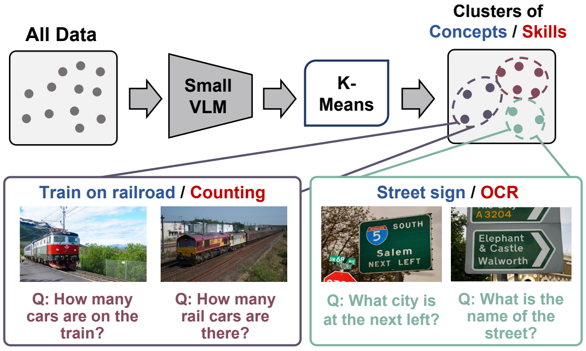

All Data
Small VLM
K-Means
Clusters of Concepts / Skills
Train on railroad / Counting
Q: How many cars are on the train?
Q: How many rail cars are there?
Street sign / OCR
SOUTH Salem NEXT LEFT
A3204
Elephant & Castle Walworth
Q: What city is at the next left?
Q: What is the name of the street?

Figure 1: Concept-skill discovery of COINCIDE. Our method utilizes a small VLM to cluster visual instruction tuning data based on concept-skill compositions.

1 Introduction

Recent studies in curriculum learning for Large Language Models (LLMs) (Chen et al., 2023; Albalak et al., 2023; Lee et al., 2024a) emphasize the importance of data ordering for effective training. However, curriculum learning for Large Vision-Language Models (LVLMs) remains unexplored. In this research, we propose leveraging vison-language (VL) concept-skill compositions identified by our recent work, COINCIDE (Lee et al., 2024b), for curriculum learning in visual instruction tuning (VIT).

Our key intuition is that LVLM abilities exist in a hierarchical structure. For the LVLM to effectively learn from VIT datasets, the model should undergo progressive training. For example, the model could effectively learn skills by following a sequential order that begins with fundamental skills like object recognition, advances to understanding object attributes and relationships, and then learns complex skills such as visual reasoning.

In our approach, we combine the concept-skill compositions and a popular reinforcement learning approach to enable LVLMs to automatically find their optimal skill learning orders for efficient and effective visual instruction tuning.

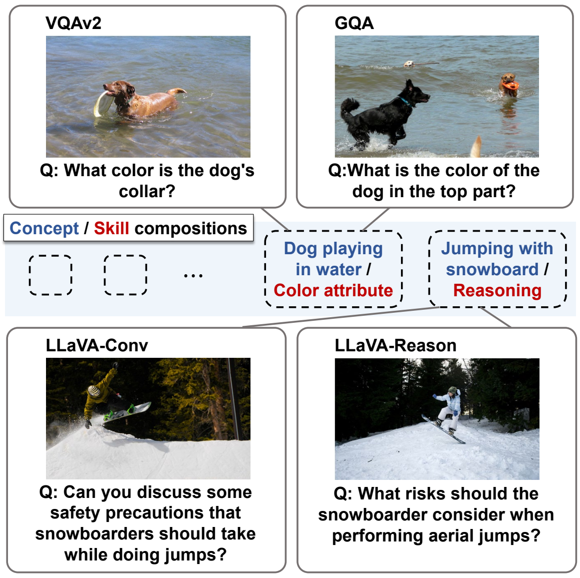

VQAv2
Q: What color is the dog's collar?
GQA
Q:What is the color of the dog in the top part?
Concept / Skill compositions
...
Dog playing in water / Color attribute
Jumping with snowboard / Reasoning
LLaVA-Conv
Q: Can you discuss some safety precautions that snowboarders should take while doing jumps?
LLaVA-Reason
Q: What risks should the snowboarder consider when performing aerial jumps?

Figure 2: VL tasks (e.g., VQAv2 and GQA, LLaVA-Conv and LLaVA-Reason) share VL concept-skill compositions.

2 Method

2.1 COINCIDE

As illustrated in Figure 1, COINCIDE uses internal neural network activations from a small model for clustering to identify VL concepts and skills in VIT data. A concept could be street signs or trains on a railroad, while a skill could be OCR, recognizing color, or reasoning.

Upon close inspection, we find that different VL tasks contain overlap over these concept-skill compositions. As exemplified in Figure 2, LLaVA-Conv and LLaVA-Reason contain questions about the risks of snowboard jumps, despite their separate focuses on multi-turn conversations and reasoning. This highlights the potential benefits of defining model skills based on shared concept clusters, offering more granular and effective skill distinctions compared to traditional task-based skill definition (e.g. GQA, VQAv2, OCRVQA) (Chen et al., 2023).

Algorithm 1 Curriculum Learning with Exp3 Algorithm

```
Require: $K$: the number of clusters, $\gamma$: exploration rate, $\mu$: moving average rate, $T$: the number of rounds $S$: the number of steps per round, $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_K\}$: Dataset for $K$ clusters, $f_\theta$: training model
Initialization: $R_i = 0$, $\mathcal{L}_{i,\text{past}} = 0 \;\; i \in \{1, 2, \dots, K\}$
for $t = 1, 2, \dots, T$ do
    $p_i(t) = (1-\gamma)\frac{R_i}{\sum_{j=1}^{K}} + \frac{\gamma}{K},\;\; i \in \{1, 2, \dots, K\}$    ▷ Calculate cluster selection distribution at step $t$
    for $s = 1, 2, \dots, S$ do
        $D_i \sim p(t)$    ▷ Select cluster $i$ from the step $t$ distribution
        $B \sim D_i$    ▷ Sample batch from cluster $i$
        $R_i \leftarrow (1-\mu)R_i + \mu(\mathcal{L}_{i,\text{past}} - \mathcal{L}(f_\theta, B))$    ▷ Update reward of cluster $i$
        $\mathcal{L}_{i,\text{past}} = \mathcal{L}(f_\theta, B)$    ▷ Update past loss value of cluster $i$
        $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(f_\theta, B)$    ▷ Update model parameters
    end for
end for
```

2.2 Skill order discovery with Exp3

We aim to design an effective curriculum learning algorithm that automatically determines the optimal training sequence using Exp3 algorithm (Allesiardo et al., 2017) and clusters derived from COINCIDE, which captures VL concept-skill compositions. Our approach begins by clustering VIT data into these concept-skill compositions using the COINCIDE method. In this context, each cluster is treated as a multi-armed bandit, where a training loss difference of a cluster is used as a reward for the Exp3 algorithm. An overview of our curriculum learning strategy is illustrated in Figure 3.

At each training step, one of the clusters is selected based on a cluster selection probability distribution $p(t)$ derived from reward signals associated with previous selections. Then, a sample batch is drawn from the chosen cluster and used to train the target model. The model subsequently receives a reward signal indicating whether the selected cluster contributed positively to the training progress, refining future cluster selections.

When a cluster shows a large training loss reduction, it indicates that the LVLM is effectively learning the skill associated with that cluster at the current stage. Consequently, the model should sample more data from this cluster to further improve training efficacy. On the other hand, if a cluster shows little to no change in training loss, it suggests that the model has either already mastered the relevant skill or is not yet in a suitable state to learn it. In such cases, the model may benefit from focusing on other clusters to ensure a more effective learning process. Detailed steps of this procedure can be found in Algorithm 1.

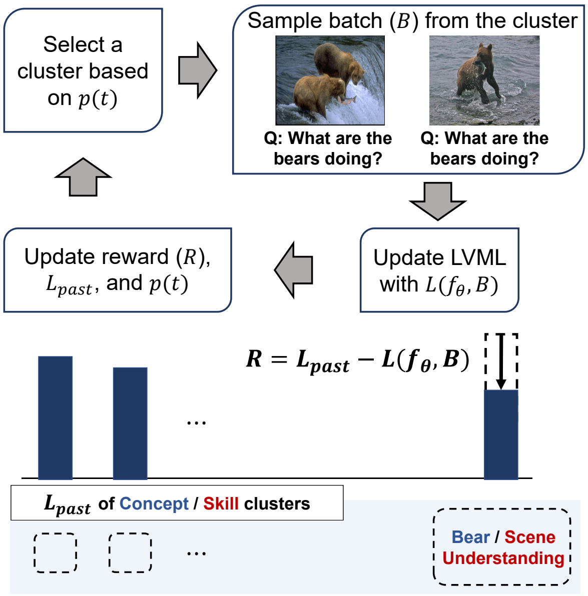

Select a cluster based on $p(t)$
Sample batch (B) from the cluster
Q: What are the bears doing?
Q: What are the bears doing?
Update LVML with $L(f_\theta, B)$
Update reward (R), L_{past}, and $p(t)$
$R = L_{past} - L(f_\theta, B)$
...
L_{past} of Concept / Skill clusters
...
Bear / Scene Understanding

Figure 3: Overview of the proposed curriculum learning pipeline. For each concept-skill composition, we track past training loss value (L_{past}) to measure its previous performance. After training an LVLM using a current batch of data, which is selected based on a probability distribution ($p(t)$), we calculate a reward (R). This reward is based on the change in the loss value, indicating whether the selected data batch improved the model's performance. Through the reward, we update $p(t)$ and repeat the above process to train the model.

3 Experiments

3.1 Setup

Visual Instruction Tuning Dataset We conduct coreset selection on LLaVA-1.5 (Liu et al., 2023a) VIT dataset. The LLaVA-1.5 dataset contains 665k VIT data from 12 different VL tasks.

Models For the target LVLMs, we use the pretrained LLaVA-1.5 model (Liu et al., 2023a) with a default size of 7B parameters. In all experiments, we train the models using LoRA (Hu et al., 2022) for one epoch, following the official finetuning hyperparameters specified in LLaVA-1.5. As a ref-

Table 1: We perform a comparison with the curriculum learning baseline on various multimodal evaluation benchmarks. We finetune LVLMs with LLaVA-1.5 (Liu et al., 2023a) dataset utilizing **20%** of the total steps for training. The best and the second best results are in **bold** and underlined, respectively.

Method	VQAv2	GQA	VizWiz	SQA-I	TextVQA	POPE	MME	MMBench		LLaVA-Bench	Rel. (%)
								en	cn		
Full-Finetune	79.1	63.0	47.8	68.4	58.2	86.4	1476.9	66.1	58.9	67.9	100
Random	75.7	**58.9**	44.3	68.5	55.3	84.7	**1483.0**	62.2	54.8	65.0	95.8
Skill-it (Chen et al., 2023)	75.4	56.9	42.4	69.7	**56.0**	84.9	1370.5	62.9	55.2	**68.2**	95.3
COINCIDE+ (Ours)	**75.8**	56.9	**49.3**	**69.9**	53.9	**86.0**	1470.0	**64.0**	**56.4**	67.6	**97.5**

erence model, we use the TinyLLaVA-2B (Zhou et al., 2024), a small VLM finetuned on the target VIT dataset, for efficient internal activation extraction and clustering. All experiments are conducted using 4 V100 GPUs.

Evaluation Benchmark To assess the generalization of finetuned LVLMs across diverse visual instructions, we evaluate the models on several widely adopted zero-shot multimodal evaluation benchmarks, including 1) visual question answering: VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018); 2) knowledge-grounded QA: ScienceQA (Lu et al., 2022); 3) Optical Character Recognition (OCR): TextVQA (Singh et al., 2019); 4) hallucination: POPE (Li et al., 2023); 5) multiple-choice: MME (Fu et al., 2023), MMBench (Liu et al., 2023c); 6) free-form generation: LLaVA-Bench (Liu et al., 2023b).

Since each evaluation benchmark has a different scale, we compute average relative performance, denoted as Rel., across benchmarks to assess the level of generalization. Each relative performance is derived from the formula: (model performance / full-finetuned performance) × 100%.

Baselines COINCIDE+ incorporates the Exp3 algorithm and VL concept-skill compositions of the COINCIDE. For comparison, we employ the Skill-it (Chen et al., 2023), a recent curriculum learning algorithm that uses task labels and evaluation loss of each task to estimate the training model's state. We additionally report the results of *Random*, the model finetuned with random sampling strategy, and *Full-Finetune*, the model finetuned with the full VIT dataset.

3.2 Results and Discussion

We conduct curriculum learning experiments on the LLaVA-1.5 dataset, utilizing 20% of the total training steps. The performance of the finetuned models is summarized in Table 1. Our results show that COINCIDE+ outperforms Skill-it by a notable 2.2 percentage points (pp) in average relative performance. Additionally, COINCIDE+ performs competitively compared to the Full-Fintune baseline, outperforming the baseline in VizWiz and SQA-I evaluation benchmarks, despite training on only 20% of the dataset.

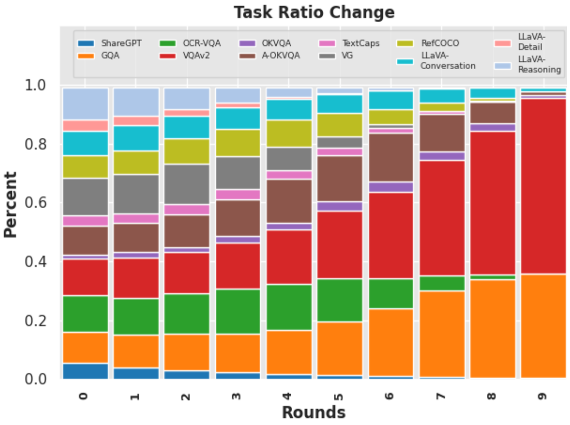


Figure 4: Visualization of task ratio for each round (t). Each task ratio is calculated as the sum of the cluster selection distribution of task-related clusters.

Interestingly, Skill-it underperforms even the Random baseline, suggesting that defining skills based solely on task names is ineffective. In contrast, COINCIDE+ achieves its improvement without relying on predefined evaluation datasets or task labels. Instead, it autonomously constructs fine-grained clusters and enables the model to dynamically select optimal data clusters during training. This approach proves both effective and applicable to real-world scenarios.

However, we observe a significant performance drop in COINCIDE+ when evaluated on the TextVQA evaluation benchmark. To investigate this, we analyze the cluster selection distribution by aggregating all the distribution values related to each task. The results, visualized in Figure 4, show that OCR-VQA-related clusters (colored green in Figure 4) are assigned a low priority in the later training rounds. Hence, we analyze that the model learned OCR skills in the middle of training, and thus the importance of these clusters would diminish over time. Consequently, the model suffers from catastrophic forgetting of the OCR abilities as it is trained on other data, yet the reward mechanism failed to capture this loss in OCR ability.

4 Future Direction

To improve our approach, we plan to extend the Exp3 algorithm to address the catastrophic forgetting of LVLM skills. Drawing from continual learning techniques, we aim to integrate a rehearsal memory mechanism that stores data for each concept-skill composition. When the model shows signs of forgetting a specific concept-skill, the stored data will be revisited during training to reinforce the model's retention of necessary concept-skill compositions.

References

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. 2023. Efficient online data mixing for language model pre-training. *arXiv preprint arXiv:2312.02406*.

Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. 2017. The non-stationary stochastic multi-armed bandit problem. *Int. J. Data Sci. Anal.*, 3(4):267–283.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2023. Skill-it! A data-driven skills framework for understanding and training language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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 International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bruce W. Lee, Hyunsoo Cho, and Kang Min Yoo. 2024a. Instruction tuning with human curriculum. In *North American Chapter of the Association for Computational Linguistics*.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024b. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tanglin 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 *Advances in Neural Information Processing Systems (NeurIPS)*.

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 International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gabriela Ben Melech Stan, Raanan Y. Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei

Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

A Visualizing LVLM Skills with Relevancy Maps

In our method, we extract neuron activations from various layers to represent the concepts and skills of each VIT data. In this approach, we hypothesize that distinct layers represent distinct concepts and skills of the LVLM. To support this assumption, we compute relevancy maps (Chefer et al., 2021) following the approach outlined in Stan et al. (2024). The relevancy maps help us understand the model's final output by highlighting the most contributing parts of the input for each layer. Given the target output token $\boldsymbol{y_t}$ and the attention map $\boldsymbol{A_l} \in \mathbb{R}^{h \times (N_v+N_l) \times (N_v+N_l)}$ of the $\boldsymbol{l}$-th layer, where h is the head dimension of the attention, the relevancy map $\boldsymbol{R}$ is computed as follows:

$$\begin{aligned} \bar{\boldsymbol{A}}_l &= \mathbb{E}_h[\nabla \boldsymbol{A}_l \odot \boldsymbol{A}_l],\ \nabla \boldsymbol{A}_l = \frac{\partial \boldsymbol{y}_t}{\partial \boldsymbol{A}_l}, \\ \boldsymbol{R} &= \boldsymbol{R} + \bar{\boldsymbol{A}}_l \cdot \boldsymbol{R}, \quad \text{for } l \in \{1, 2, \ldots, L\}, \end{aligned} \tag{1}$$

where $\odot$ denotes the Hadamard product and L is the total number of layers in the LVLM. In order to investigate the contribution of each layer to the final output, we visualize the image regions related to the output token through the visual relevancy map computed from each layer. Specifically, we consider the row of $\bar{\boldsymbol{A}}_l \cdot \boldsymbol{R}$ corresponding to the output token. Then, we extract the visual token parts of the row to yield the visual relevancy map.

For the investigation, we inspect the 4th, 8th, 12th, 16th, and 20th layers of the TinyLLaVA-2B (Zhou et al., 2024) model and identify the layer that activates the most relevant visual regions. The results, shown in Figure 5, reveal that (1) the most relevant layer varies according to the concept-skill composition and (2) the most relevant layer is the same across diverse VIT data when the data shares a similar concept-skill composition. These findings support our initial assumption that different layers contribute to distinct concepts and skills. Therefore, using neuron activations from diverse layers can effectively group VIT data according to their concept-skill composition.

B Concept-Skill Clustering Visualization

We visualize the clustering results of the gathered VIT data. The results are illustrated in Figure 6. We observe that most clusters contain VIT data that encode similar concept-skill compositions. For instance, the first group in Figure 6 consists of samples requiring OCR and counting abilities to solve visual queries involving images with store signs. The second group features images of people waiting for public transportation and multiple-choice questions that require visual recognition and reasoning abilities. The third group shows a cluster of samples with images of people in suits and queries focusing on object localization and generating captions for given bounding boxes. Lastly, the bottom group includes images exhibiting children with animals and requiring the ability to reason about the educational benefits that the children might gain from interacting with the animals.

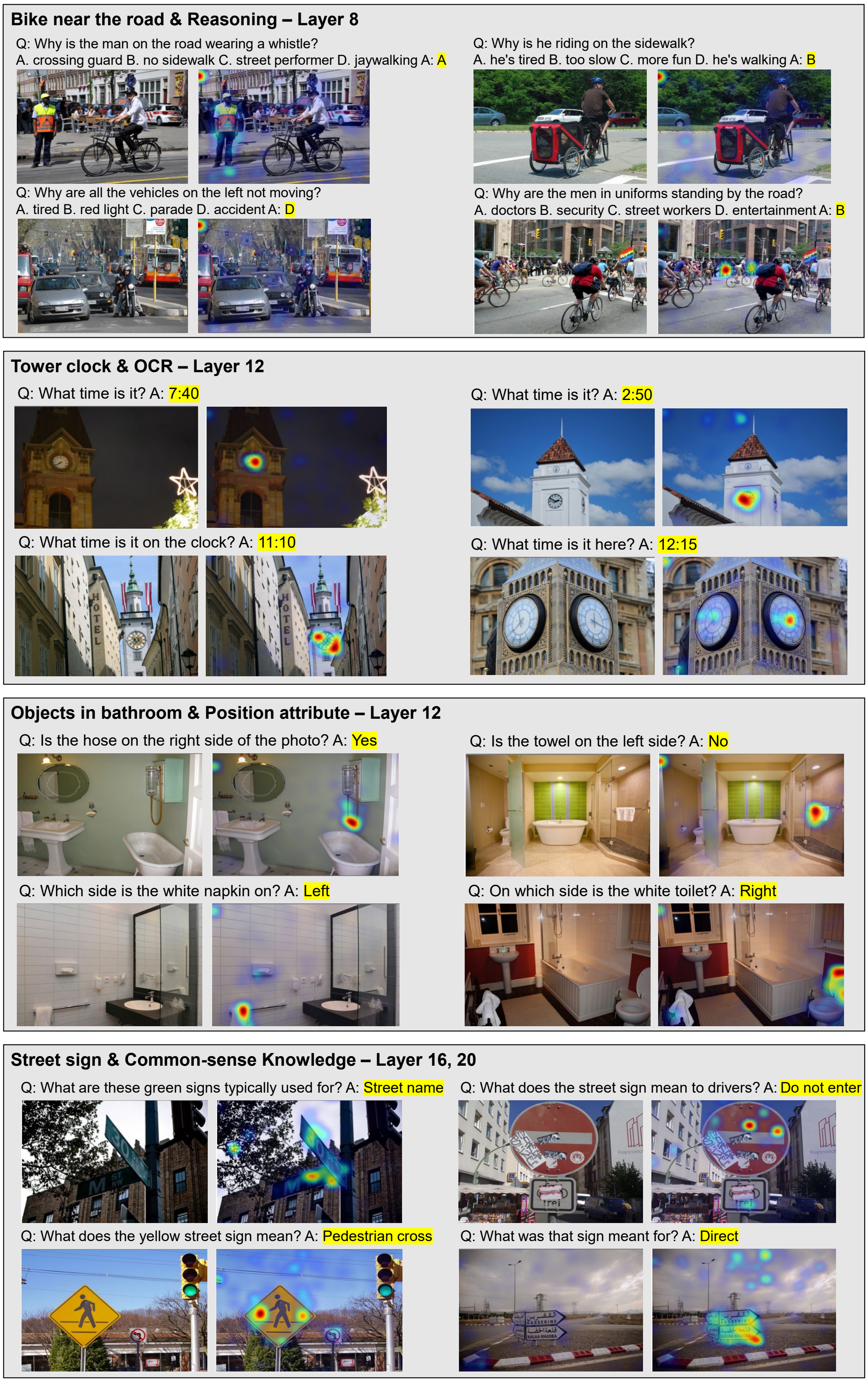
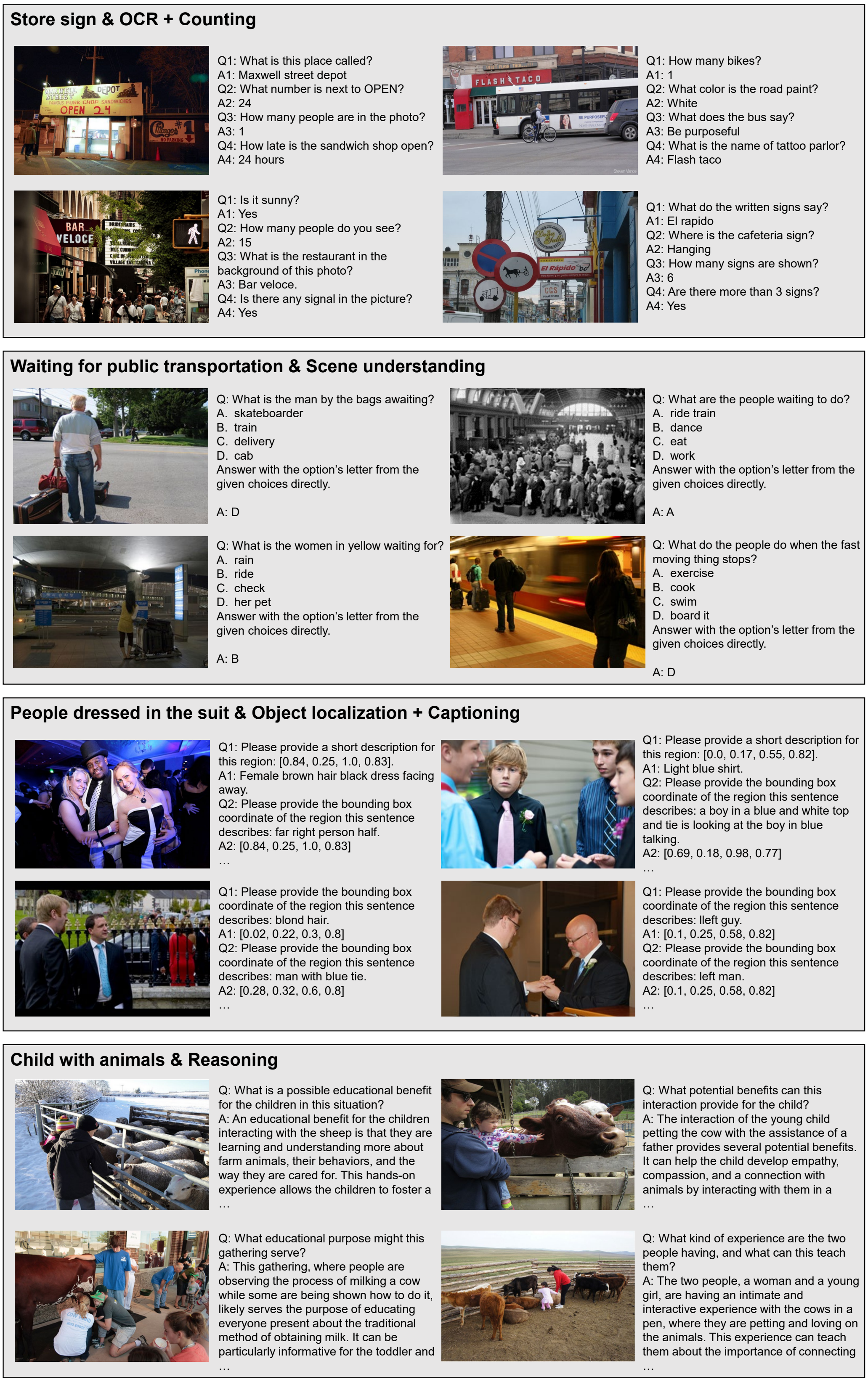

Bike near the road & Reasoning – Layer 8
Q: Why is the man on the road wearing a whistle?
A. crossing guard B. no sidewalk C. street performer D. jaywalking A: A
Q: Why is he riding on the sidewalk?
A. he's tired B. too slow C. more fun D. he's walking A: B
Q: Why are all the vehicles on the left not moving?
A. tired B. red light C. parade D. accident A: D
Q: Why are the men in uniforms standing by the road?
A. doctors B. security C. street workers D. entertainment A: B
Tower clock & OCR – Layer 12
Q: What time is it? A: 7:40
Q: What time is it? A: 2:50
Q: What time is it on the clock? A: 11:10
Q: What time is it here? A: 12:15
Objects in bathroom & Position attribute – Layer 12
Q: Is the hose on the right side of the photo? A: Yes
Q: Is the towel on the left side? A: No
Q: Which side is the white napkin on? A: Left
Q: On which side is the white toilet? A: Right
Street sign & Common-sense Knowledge – Layer 16, 20
Q: What are these green signs typically used for? A: Street name
Q: What does the street sign mean to drivers? A: Do not enter
Q: What does the yellow street sign mean? A: Pedestrian cross
Q: What was that sign meant for? A: Direct

Figure 5: Relevancy maps visualization. We investigate which layer contributes most to the final output of the LVLM. This is done by visualizing relevancy maps of four samples from the same cluster. For each example, the left image is the original, while the right image shows the visualized relevancy map, highlighting regions most relevant to the LVLM output text colored in yellow. The top-left corner of each group explains the VL concept-skill composition and the layer number with the highest relevancy to the output.



Store sign & OCR + Counting

Q1: What is this place called?
A1: Maxwell street depot
Q2: What number is next to OPEN?
A2: 24
Q3: How many people are in the photo?
A3: 1
Q4: How late is the sandwich shop open?
A4: 24 hours

Q1: How many bikes?
A1: 1
Q2: What color is the road paint?
A2: White
Q3: What does the bus say?
A3: Be purposeful
Q4: What is the name of tattoo parlor?
A4: Flash taco

Q1: Is it sunny?
A1: Yes
Q2: How many people do you see?
A2: 15
Q3: What is the restaurant in the background of this photo?
A3: Bar veloce.
Q4: Is there any signal in the picture?
A4: Yes

Q1: What do the written signs say?
A1: El rapido
Q2: Where is the cafeteria sign?
A2: Hanging
Q3: How many signs are shown?
A3: 6
Q4: Are there more than 3 signs?
A4: Yes

Waiting for public transportation & Scene understanding

Q: What is the man by the bags awaiting?
A. skateboarder
B. train
C. delivery
D. cab
Answer with the option's letter from the given choices directly.

A: D

Q: What are the people waiting to do?
A. ride train
B. dance
C. eat
D. work
Answer with the option's letter from the given choices directly.

A: A

Q: What is the women in yellow waiting for?
A. rain
B. ride
C. check
D. her pet
Answer with the option's letter from the given choices directly.

A: B

Q: What do the people do when the fast moving thing stops?
A. exercise
B. cook
C. swim
D. board it
Answer with the option's letter from the given choices directly.

A: D

People dressed in the suit & Object localization + Captioning

Q1: Please provide a short description for this region: [0.84, 0.25, 1.0, 0.83].
A1: Female brown hair black dress facing away.
Q2: Please provide the bounding box coordinate of the region this sentence describes: far right person half.
A2: [0.84, 0.25, 1.0, 0.83]
…

Q1: Please provide a short description for this region: [0.0, 0.17, 0.55, 0.82].
A1: Light blue shirt.
Q2: Please provide the bounding box coordinate of the region this sentence describes: a boy in a blue and white top and tie is looking at the boy in blue talking.
A2: [0.69, 0.18, 0.98, 0.77]
…

Q1: Please provide the bounding box coordinate of the region this sentence describes: blond hair.
A1: [0.02, 0.22, 0.3, 0.8]
Q2: Please provide the bounding box coordinate of the region this sentence describes: man with blue tie.
A2: [0.28, 0.32, 0.6, 0.8]
…

Q1: Please provide the bounding box coordinate of the region this sentence describes: lleft guy.
A1: [0.1, 0.25, 0.58, 0.82]
Q2: Please provide the bounding box coordinate of the region this sentence describes: left man.
A2: [0.1, 0.25, 0.58, 0.82]
…

Child with animals & Reasoning

Q: What is a possible educational benefit for the children in this situation?
A: An educational benefit for the children interacting with the sheep is that they are learning and understanding more about farm animals, their behaviors, and the way they are cared for. This hands-on experience allows the children to foster a
…

Q: What potential benefits can this interaction provide for the child?
A: The interaction of the young child petting the cow with the assistance of a father provides several potential benefits. It can help the child develop empathy, compassion, and a connection with animals by interacting with them in a
…

Q: What educational purpose might this gathering serve?
A: This gathering, where people are observing the process of milking a cow while some are being shown how to do it, likely serves the purpose of educating everyone present about the traditional method of obtaining milk. It can be particularly informative for the toddler and
…

Q: What kind of experience are the two people having, and what can this teach them?
A: The two people, a woman and a young girl, are having an intimate and interactive experience with the cows in a pen, where they are petting and loving on the animals. This experience can teach them about the importance of connecting
…

Figure 6: Examples of data clusters. We visualize four samples from the same cluster. The top-left corner of each group explains the VL concept-skill composition.