Beyond Sight: Towards Cognitive Alignment in LVLM via Enriched Visual Knowledge

Supplementary Material

A. Experiment Details

Architecture. We utilize CLIP-ViT-L-14 [29] as the vision encoder, with a default resolution of 336×336 , and Meta-Llama-3-8B [1] as the language model. The low-resolution (LR) branch employs a 2-layer MLP as the adapter, while the high-resolution (HR) branch compresses visual tokens using a shared Perceiver resampler layer [2]. In the HR branch, each high-resolution image is divided into four sub-patches. The resampler processes the visual tokens from these sub-images, compressing them from 2,880 down to 128 tokens via cross-attention with query vectors. These 128 tokens are then concatenated with the 576 tokens from the low-resolution overview image and fed into the LLM.

Pretrain Datasets. We use the same dataset for LLaVA-1.5 experiments. Specifically, stage 1 uses CC595k [31] and stage 2 uses DataMix 665k [10, 11, 13, 14, 20, 23–25, 30, 33] proposed in Liu et al. [19].

Hyperparameters. In this work, we adopt the same set of hyperparameters as LLaVA-1.5 [19]. We show the training hyperparameters for LLaVA-1.5 experiments in Table 5. All experiments are conducted using a maximum of 8 Nvidia H800 GPUs. We set the value of $\mu_e = 7.32$ and $\mu_h = 4.38$, respectively, with a sparsification threshold $\theta = 0.5$ applied to selectively filter out lower relevance tokens. Additionally, the temperature parameter, initialized at zero, is set as learnable to dynamically adjust throughout training.

Robustness of EECA. We evaluate the robustness of EECA across different hyperparameter settings. The model maintains comparable accuracy within specific ranges of key parameters, such as the balancing coefficient μ_e , sparsification threshold θ , and high-resolution visual tokens N_{v_H} (See Figure 7), demonstrating stable performance despite small variations.



Figure 7. The robustness of the different hyperparameters.

B. Datasets Construction for MGLD

This section expands Section 3 with additional details about our data preprocessing steps.

Uuparparamatar	LLaVA-1.5		EECA	
Hyperparameter	Stage 1	Stage 2	Stage 3	
batch size	256	128	128	
lr	2e-3	2e-5	2e-5	
lr schedule decay	cosine	cosine	cosine	
lr warmup ratio	0.03	0.03	0.03	
weight decay	0	0	0	
epoch	1	1	1^{*}	
optimizer	AdamW [21]			
DeepSpeed stage	2	2	2	

Table 5. Hyperparameters for EECA training on LLaVA-1.5. By default, EECA is trained for 1 epoch, denoted by *. Unless otherwise stated, the results presented in Table 4 are based on 2 epochs of training.

B.1. Details of google landmarks dataset v2

Overview. The Google Landmarks Dataset v2 (GLDv2) [41] is the largest benchmark for fine-grained instance recognition and image retrieval, comprising over 5 million images with 200,000 distinct instance labels. It sourced from Wikimedia Commons, and is characterized by real-world challenges such as imbalanced class distribution and high intra-class variability.

Data usage in this work. In this study. we leverage the GLDv2 dataset to construct a finetuning dataset, utilizing all data from the training set (train.csv, train_label_to_category.csv, and This dataset train_label_to_hierarchical.csv). comprises 4.1 million images spanning 203,000 landmarks.

- train.csv: Contains fields id, url, and landmark_id. Here, id is a 16-character string, url is a string representing the image's URL, and landmark_id is an integer identifier for the landmark.
- train_label_to_category.csv: Includes landmark_id and category fields. landmark_id is an integer, while category is a Wikimedia URL linking to the class definition of the landmark.
- train_label_to_hierarchical.csv: Contains fields landmark_id, category, supercategory, hierarchical_label, and natural_or_human_made. Supercategory refers to the type of landmark (e.g., natural or human-made), mined from Wikimedia. Hierarchical_label corresponds to the landmark's hierarchical classification, and natural_or_human_made indicates whether the landmark is naturally occurring or man-made.

The structure of the GLDv2 training dataset is depicted in Figure 8. Each hierarchical_label encompasses multiple categories,



Figure 8. The structure of the GLDv2 train set

Figure 9. Frequency of the counts of images per category

and each category consists of a varying number of images, reflecting the diverse and hierarchical nature of the dataset. The category distribution in GLDv2 training dataset is highly imbalanced, as illustrated in Figure 9. Approximately 57% of the categories contain at most 10 images, and 38% have 5 or fewer images. This makes the dataset diverse, covering a wide range of landmarks, from globally renowned sites to more obscure, local landmarks.

B.2. Image selection methodology

In the first stage (Figure 2), we select representative images from the GLDv2 dataset. Since each landmark name corresponds to multiple images, typically sourced from Wikipedia entries related to the landmark, we use the landmark's simple name (*e.g.* "Eiffel Tower") as the sole reference. Using CLIP-based similarity measures, we select the image that best matches this name, filtering out low-quality or ambiguous photos and ensuring a high-quality visual representation aligned with the landmark's identity. From the top three images with the highest similarity scores, we conduct weighted sampling based on their similarity to select a unique image corresponding to each landmark. Figure 10 presents a specific example of image selection.

B.3. Prompt design for data annotation

Q-A pair. Once we have a refined set of images, we generate Question-Answer (Q-A) pairs to facilitate landmark recognition. The prompt used for image Q-A pair annotation is shown in Figure 12. For each selected image, we random select one question

Figure 10. Example for the image selection methodology (the Castle of Pardo de Cela). **Red**: The top3 images with the highest similarity scores. **Green**: The final image obtained through weighted sampling.

from the questions set and add the landmark name as a reference to ensure the accuracy of the annotation. While answering these questions, the model is encouraged to provide descriptive details about the landmark, drawing on both the visual features and contextual information. This approach aims to broaden the model's understanding of each landmark's visual and contextual uniqueness, laying a foundation for aligning with VE's cognitive space.

Multi-granularity data annotation. Following the Q-A pair annotation for each image, we generate multi-granularity data using the multi-granularity data generation prompt shown in Figure 11. In the original dataset, some hierarchical labels were already provided. To enhance accuracy, we employed GPT-40 to refine and expand the annotations, using the original labels as a reference. This process provides us with entities and an updated hierarchical label.

Figure 11. Multi-Granularity data generation prompt. The description is the Q-A pair.

Figure 12. Q-A pair prompt.

Figure 13. Example of the MGLD datasets. The conversation is the Q-A pair, and the hierarchical_label_gpt4o is the new hierarchical label annotated by GPT-4o.

B.4. MGLD overview.

We structure the data as illustrated in Figure 13. Each image is associated with a Q-A conversation, its landmark name, entities that capture both visual and conceptual features, and a hierarchical label representing its general category.

From the final dataset of approximately 203k samples, we set aside 5k samples as the test dataset.

C. Evaluation Detail

id": 42349

C.1. Prompt design

We provide GPT40 with 5 inference runs of the VLLM and the ground-truth answer. The GPT40 is asked to evaluate the overall recognition of the landmark based on all 5 responses together, and classify the level of recognition into one of the four levels: Strongly Known, Known, Partially Known, or Unknown. The classification criteria is clearly defined in the prompt (See Figure 14).

Figure 14. Evaluation prompt.

C.2. Rating criteria examples

This section provides an example ("Kinderdijk Windmills") evaluated by GPT-40, where the answer across different models is assessed at four different levels—*Strongly Known, Known, Partially Known, and Unknown.* Figure 15 shows the specific answers from each model, followed by GPT-40's explanations for the corresponding evaluations.

- Explanation of strongly known : The responses correctly identify the Kinderdijk Windmills multiple times, specifically in Responses 2 and 3, providing detailed and accurate information that matches the correct answer. These responses include precise details about the location, historical significance, and characteristics of the Kinderdijk Windmills, demonstrating a strong recognition of the landmark.
- Explanation of known: Response 1 **correctly identifies the Kinderdijk Windmills** and provides detailed and accurate information about their location, history, and significance. This response alone is sufficient to classify the recognition as 'Known'. Although Responses 2 and 3 incorrectly identify the location as Zaanse Schans, they do **provide relevant details about Dutch windmills**, which are related to the correct landmark. Responses 4 and 5 describe general features of Dutch windmills and mention regions like Kinderdijk, but do not specifically identify the landmark. Overall, the presence of one accurate identification with detailed information justifies the 'Known' classification.
- Explanation of weakly unknown: None of the responses correctly identify the 'Kinderdijk Windmills' specifically. However, several responses mention relevant Dutch windmills and

Figure 15. Model responses from different recognition level. There is a significant gap in recognition ability among "Strongly Known," "Weakly Unknown," and "Unknown". The difference between "Strongly Known" and "Known" is the number of times the model correctly identifies the landmark.

regions associated with windmills, such as Kinderdijk and Zaanse Schans, providing geographical hints and cultural context related to Dutch windmills. This suggests a connection to the correct landmark, even though it is not explicitly named.

• Explanation of unknown: None of the responses correctly identify the 'Kinderdijk Windmills'. The responses mention other windmills and locations such as 'Weerschipmolen', 'Zaanse Schans', and 'Huis op de Hoop', but these are not related to the correct landmark. The information provided is either unrelated or significantly off the mark, as none of the responses provide any specific details or hints that connect to the Kinderdijk Windmills.

D. Statistical Analysis for Partial Knowledge Parts

D.1. Detail results for ablation study

We reports the results of the ablation study in terms of proportions relative to the entire dataset (See Table 3). Here, in Table 6, we provide the corresponding absolute counts for each knowledge level (Strongly Known, Known, Weakly Unknown, Unknown).

Method	Strongly Known	Known	Weakly Unknown	Unknown
Baseline	103	114	145	2138
+ HSS-50k	187	161	145	2007
+ HR-Branch	198	149	152	2001
+ \mathcal{L}_e	212	148	163	1977
+ \mathcal{L}_h	213	175	159	1953

Table 6. Detailed ablation study results. Absolute counts for each level of knowledge.

D.2. Detailed results for generalizability experiment

We summarizes the results of the generalizability experiments using accuracy (See Table 4). To complement this, Table 7 presents the absolute counts for each knowledge level (Strongly Known, Known, Weakly Unknown, and Unknown).

Method	Strongly Kn	own Known W	eakly Unknov	wn Unknown	
Baseline	103	114	145	2138	
		For HDS-25k			
+Data	182	142	151	2025	
+HR Branch	218	122	185	1975	
$+\mathcal{L}_h$	233	127	175	1965	
$+\mathcal{L}_e$	229	121	199	1951	
For HSS-25k					
+Data	169	135	201	1995	
+HR Branch	202	144	177	1977	
$+\mathcal{L}_h$	205	141	193	1961	
$+\mathcal{L}_e$	226	134	153	1987	
For LCS-25k					
+Data	179	88	253	1980	
+HR Branch	203	99	243	1955	
$+\mathcal{L}_h$	192	116	254	1938	
$+\mathcal{L}_e$	198	110	266	1926	

Table 7. Generalizability experiment results. Absolute counts for each knowledge level across datasets and methods.

E. Generalization

E.1. Generalization beyond landmarks

In this section, we provide a detailed evaluation of EECA's generalization capabilities beyond the landmark recognition, specifically assessing its performance on the iNaturalist2021 dataset [38]. This dataset comprises over 10,000 species, spanning diverse categories such as plants, insects, birds, and various other organisms, thus providing a robust benchmark for testing generalization.

Category	Reference (BRS)	VE-Unknown (LCS)	VE-Known (HDS)
Increase (%)	36.36	36.36	70.91

Table 8. Accuracy increase across different data selection methods. The baseline method (BRS) serves as a reference, while VE-Known methods (HDS), show significant improvements over both BRS and VE-Unknown (LCS), highlighting the advantage of leveraging known visual entities.

	Baseline	+ Data	+ HR Branch	+ \mathcal{L}_e
Accuracy	5.5	10.9 (+5.4)	11.4 (+0.5)	12.5 (+1.1)

Table 9. Ablation study results demonstrating the incremental effectiveness of EECA components. The integration of additional data, high resolution branch, and $\mathcal{L}e$ each yields measurable performance gains.

Our experiments demonstrate that cognitive misalignment remains a significant challenge across diverse visual recognition tasks, not limited to landmarks alone. Figure 16 highlights specific instances where standard visual models struggle to differentiate visually similar species, underlining persistent cognitive misalignment issues. Employing EECA's annotation and data selection methodologies on the iNaturalist dataset, we performed comprehensive data selection experiments Table 8 and ablation studies Table 9. Results, as summarized in the provided tables, confirm that EECA's data selection strategy significantly improves model accuracy. These findings suggest that EECA effectively mitigates cognitive misalignment in complex, real-world classification scenarios beyond landmarks.

E.2. Scalability of entity-level annotation

Entity attributes, whether pertaining to biological species, landmarks, or artifacts, are inherently definable and identifiable across multiple domains. Large language models (LLMs) pre-trained on extensive datasets such as Wikipedia, Common Crawl, and domain-specific corpora possess rich knowledge, which naturally supports scalable and cost-effective entity-level annotation. To validate scalability, we also explored open-source LLMs, which successfully generated discriminative and precise entity annotations (See Figure 16), further reinforcing the approach's broad applicability.

E.3. Generalization across architectures

Cognitive misalignment presents a fundamental limitation in large vision-language models (LVLMs), representing a previously overlooked source of errors. This issue arises inherently from the modular design of LVLMs, where the vision encoder, projector, and LLM components struggle to align representations across modalities. Importantly, Table 10 reveals that this misalignment persists regardless of the specific vision encoder (e.g., SigLIP, CLIP) or

Figure 16. Left: Illustration of cognitive misalignment where standard visual models struggle to differentiate visually similar species, exemplified by the case of *Culaea inconstans*. Right: Even smaller open-source LLMs can generate precise and discriminative entity annotations, demonstrating the scalability and cost-effectiveness of automated labeling across domains.

language model (e.g., Vicuna, LLaMA), indicating that it is a universal challenge rather than an architecture-specific bottleneck.

Vision Encoder(VE)	LLM	LCS	BRS	LDS
SigLIP-ViT	Vicuna-7B	16.16	15.96	17.52
CLIP-ViT	LLaMA3-8B	10.68	12.16	12.96

Table 10. Performance comparison of LVLMs with different vision encoders and language models.