Generative Vision-Language Multiple Instance Learning for Weakly Supervised Neonatal Fundus Screening and Reporting

Supplementary Material

Table 8. Statistics of the Instance-Level Annotated Subset of the NFSD Dataset.

Label	Image	Train	Valid	Test	Sum
RH	Total	831	282	284	1,397
	Normal	134	50	51	235
ROP	Total	1,676	557	553	2,786
	Normal	979	325	321	1,625
WS	Total	605	195	197	997
	Normal	325	102	105	532

6. NFSD Dataset Statistics

As is described in Sec. 4.1, we curated a subset of images with instance-level clinical descriptions for multi-modal learning. This annotated subset comprises 5,180 images from 526 neonates. Each image is accompanied by a textual description provided by experienced ophthalmologists, detailing the observed retinal findings.

Tab. 8 summarizes the distribution of annotated instances across the three diagnostic categories—*Retinal Hemorrhage* (RH), ROP, and *White Spots* (WS)—as well as the number of images labeled as *Normal* (images shows no obvious abnormality) within each category. The annotations are proportionally distributed across the training, validation, and test splits, mirroring the partitioning of the full dataset.

7. VLM Architecture

LLaVA-OneVision [15] follows the standard LLaVA architecture [23], comprising a vision encoder, a multimodal projection module, and an LLM. In our work, we employ `llava-onevision-qwen2-7b-si` as the description generator. This model integrates the SigLip-SO400M vision encoder [50] with the Qwen2-7B-Instruct LLM [48], whose hidden dimensions are 1152 and 3584, respectively. The multimodal projector is a two-layer MLP with a GELU activation, responsible for aligning visual features from the vision encoder with the language model's embedding space.

During fine-tuning, both the vision encoder and LLM backbone are frozen, while only the projection module and LoRA parameters are updated. The detailed fine-tuning configuration is provided in Tab. 9.

Table 9. Fine-tuning hyperparameters for LLaVA-OneVision.

Category	Setting
LoRA Configuration	
LoRA rank (r)	32
LoRA alpha	128
Training Setup	
Epochs	3
Batch size	1
Gradient accumulation steps	8
Learning rate	1×10^{-4}
Warmup ratio	0.03
LR scheduler	Cosine decay
Weight decay	0
Precision	bfloat16
Gradient checkpointing	Yes
Data / Vision Settings	
Image aspect ratio	AnyRes
Grid pinpoints	$(1 \times 1) \rightarrow (3 \times 3)$
Patch merge type	Spatial unpad

8. Compared Methods

8.1. Descriptions

In this section, we briefly summarize the compared MIL methods:

- **ABMIL** [12] learns instance importance via an attention mechanism and aggregates instances into a bag-level representation for classification.
- **CLAM** [24] combines attention-based key instance selection with instance-level classifiers trained on high- and low-attention samples, improving bag-level prediction through hybrid supervision.
- **DFTD** [51] forms pseudo-bags by grouping instances and applies attention within each group, using instance-level distillation to enhance discriminative bag-level classification.
- **DSMIL** [14] adopts a dual-stream architecture that fuses instance-level predictions with attention-derived bag features for robust MIL performance.
- **ILRA** [45] employs low-rank attention blocks to compress instance features into a latent space and reconstruct them, enabling efficient global feature aggregation.
- **R^{2}T** [38] introduces region-based recursive transformers with multi-scale processing to hierarchically aggregate instance features for MIL.

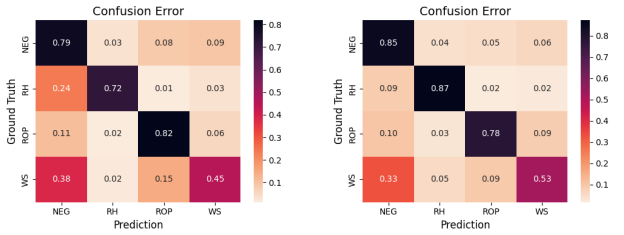
Table 10. Learning rate for every compared methods, each

MIL	Learning Rate	Epoch
ABMIL	1×10^{-4}	200
CLAM	1×10^{-5}	200
DFTD	2×10^{-5}	200
DSMIL	1×10^{-4}	200
ILRA	6×10^{-5}	200
R^2T	6×10^{-5}	200
TransMIL	5×10^{-6}	200
Transformer	5×10^{-6}	200
WIKG	1×10^{-4}	200
LD2GMIL	4×10^{-5}	25
ViLa-MIL	1×10^{-6}	100
GVL-MIL(ours)	2×10^{-5}	20

- **Transformer** [40] applies standard multi-head self-attention to model full pairwise instance dependencies, followed by CLS-token aggregation without MIL-specific architectural modifications.
- **TransMIL** [32] utilizes Nyström attention for long-sequence efficiency and square padding for irregular instance counts, focusing on CLS-to-instance relations for MIL classification.
- **WIKG** [18] integrates top-k attention with graph-inspired bi-interaction aggregation to model instance relations more effectively.
- **LD2GMIL** [7] performs instance re-embedding via self-attention and leverages a bag-prior loss using bag labels and instance weights. We adopt LD2GMIL on top of DSMIL.
- **ViLa-MIL** [33] adapts vision-language models to MIL by dual-scale descriptive prompts, a prototype-guided patch decoder, and a context-guided text decoder for multi-granular reasoning.

8.2. MIL Training Argumentations

We adopt a global batch size of 64 (computed as per-device batch size multiplied by gradient accumulation steps) for all compared MIL methods. Method-specific learning rates are listed in Tab. 10. All remaining training hyper-parameters are kept consistent across methods, including 25% negative-sample resampling, a warmup ratio of 0.05, weight decay of 0.01, and a dropout rate of 0.25. We employ early stopping with a patience of 10 epochs; all models converged well before this limit and demonstrated stable performance plateaus.



(a) LD2G-MIL. (b) GVL-MIL (ours).

Figure 6. Confusion matrix of LD2G-MIL and GVL-MIL.

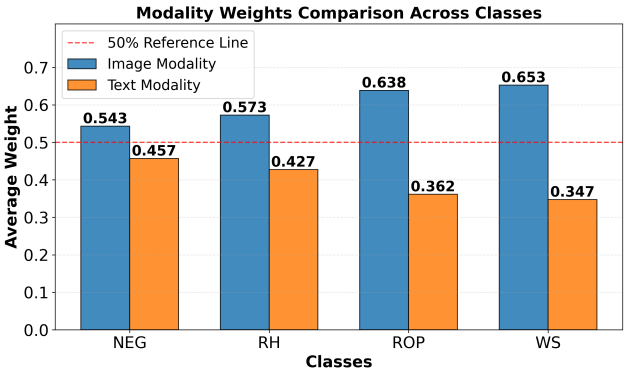


Figure 7. Modality weights across 4 classes.

9. Visualization

9.1. Description Generation

We select one annotated case from each of the ROP, RH, and WS categories and present two representative examples, as shown in Tab. 11. The second and third columns report the clinically annotated descriptions and the corresponding outputs generated by the VLM, respectively.

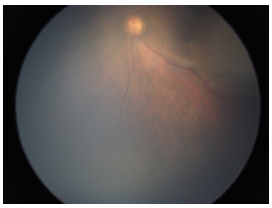
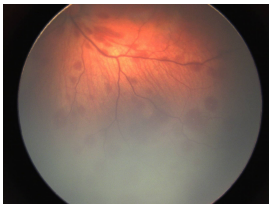
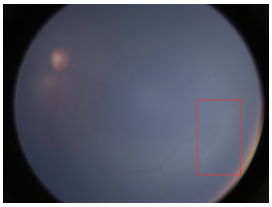
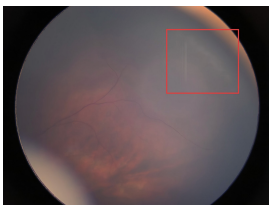
9.2. Confusion Matrix

We compare final checkpoints of LD2G-MIL and GVL-MIL in Fig. 6. LD2G-MIL shows higher misclassification across categories, whereas GVL-MIL exhibits stronger diagonal elements and weaker off-diagonals, indicating improved accuracy and reduced errors across all sample types.

9.3. Entropy Weights Between Modalities

We report the modality weights for the image and text branches (*i.e.*, w_V and w_T in Eq. (11)) across four diagnostic categories (grouped by ground truth). The mean values for each category are presented in Fig. 7. For samples diagnosis as *Negative* or *RH*, the weights of the two modalities are relatively close. In contrast, for positive cases of ROP and WS, the weight of the image modality is consistently much higher than that of the text modality. One possible reason for this phenomenon is that the lesions in ROP and WS samples are relatively small, and there is a higher

Table 11. Example of NFSD Dataset with Automatic Line Wrapping

Image ID	Expert Annotation (Translated)	Model Generation (Translated)	Bag Label
	No obvious abnormality	No obvious abnormality	NEG
	Shallow intraretinal hemorrhages arranged along the course of the retinal nerve fibers are observed in the inferotemporal retina, appearing red in color.	Shallow intraretinal hemorrhages arranged along the course of the retinal nerve fibers are observed in the nasal retina, appearing red in color, with some lesions showing yellow-white deposits at the center.	RH
	A gray-white demarcation line is visible in the nasal peripheral retina, with an avascular zone anterior to the line and markedly tortuous and dilated vessels posterior to it.	A gray-white demarcation line is visible in the temporal peripheral retina, with an avascular zone anterior to the line.	ROP
	Scattered gray-white, irregularly shaped exudative lesions with well-defined borders and no elevation are seen in the superotemporal peripheral retina.	Gray-white, irregularly shaped exudative lesions with well-defined borders and no elevation are seen adjacent to the vessels in the superotemporal peripheral retina.	WS

proportion of normal samples. As a result, the descriptions generated by the VLM tend to be of lower quality, making the model rely more heavily on information from the image modality.

9.4. Instance Weights in GVL-MIL

We visualize the instance-level attention weights for one representative bag from each class, as shown in Figs. 8 to 11. The softmax-normalized attention tends to collapse toward nearly one-hot distributions; therefore, to more clearly reveal variation across instances, we present the raw attention values after L1 normalization.

Despite the apparent softmax collapse, GVL-MIL performs strongly on the NFSD dataset because it produces discriminative instance representations and accurately identifies the most informative instances. As illustrated in the figures, the attentions from V-MIL (image modality) exhibit substantial variation across instances, whereas those from T-MIL (text modality) appear more uniform. We attribute this difference to their architectural roles: the image branch employs a dedicated aggregator that explicitly models cross-instance differences, resulting in larger spread in attention scores. In contrast, the text branch directly extracts class-relevant semantic cues for classification, leading to a more even distribution. Nevertheless, this complementary behavior enhances the stability and robustness of the overall model prediction.

9.5. Attention Map Aggregator

To demonstrate the effectiveness of the Aggregator, we select several bags and visualize the instances with the highest attention weights within each bag, as shown in Figs. 12 to 15. To more clearly illustrate the model's focus on lesion regions, we overlay the Aggregator attention maps on the original images. For ROP and WS cases, where the lesions are relatively subtle, we additionally indicate their approximate locations on the original images to further validate the attention. The visualized instances correspond closely with the associated textual descriptions, highlighting the effectiveness of the proposed module.

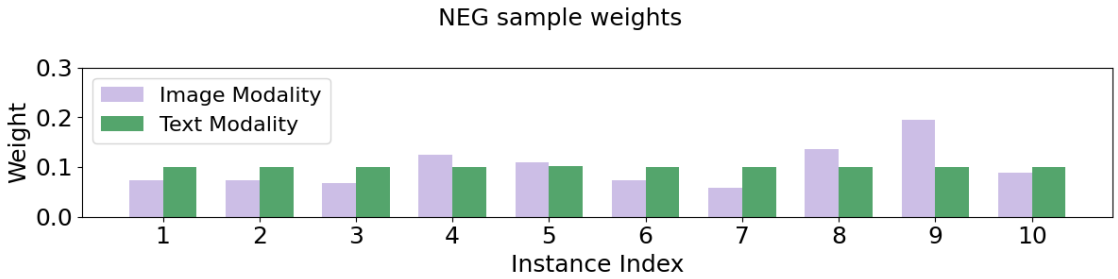


Figure 8. Instance weights of a negative sample.

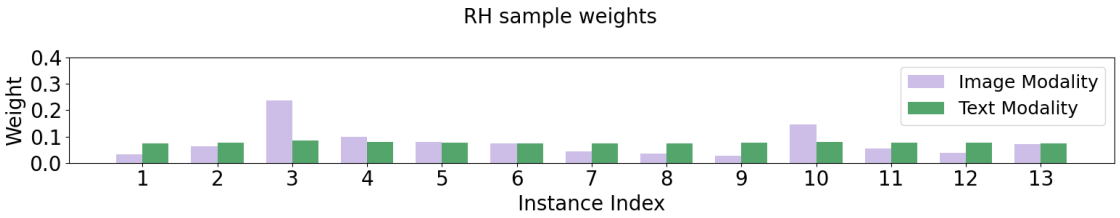


Figure 9. Instance weights of a sample with the label of RH.

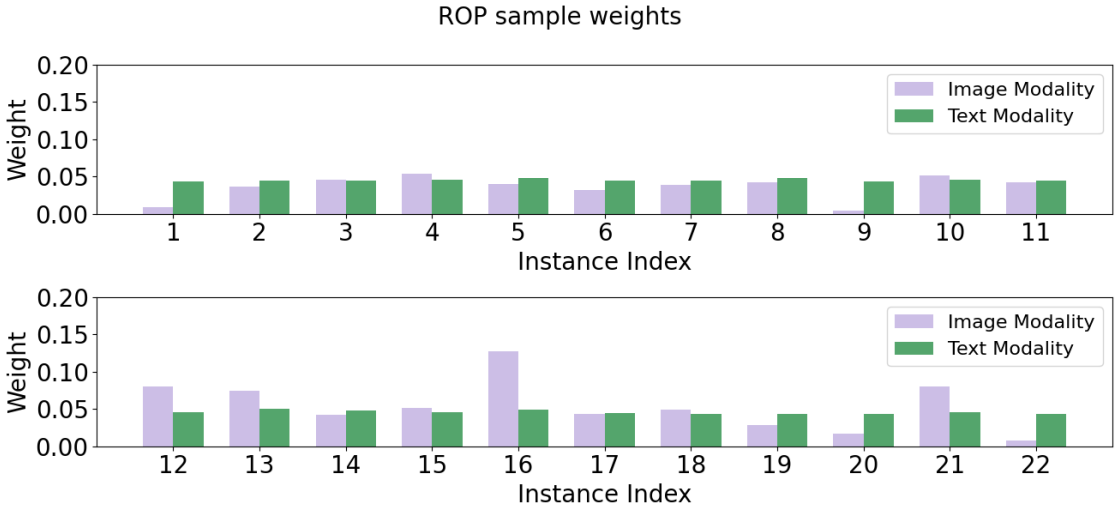


Figure 10. Instance weights of a sample with the label of ROP.

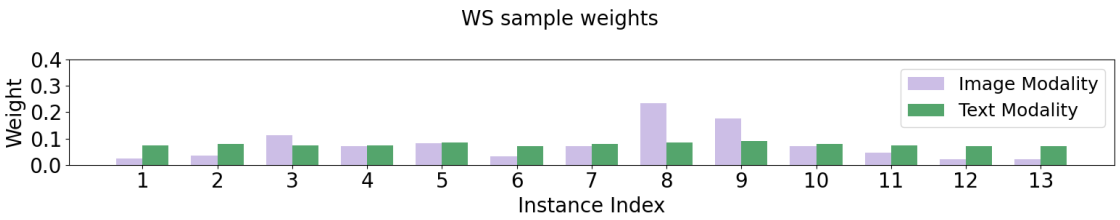
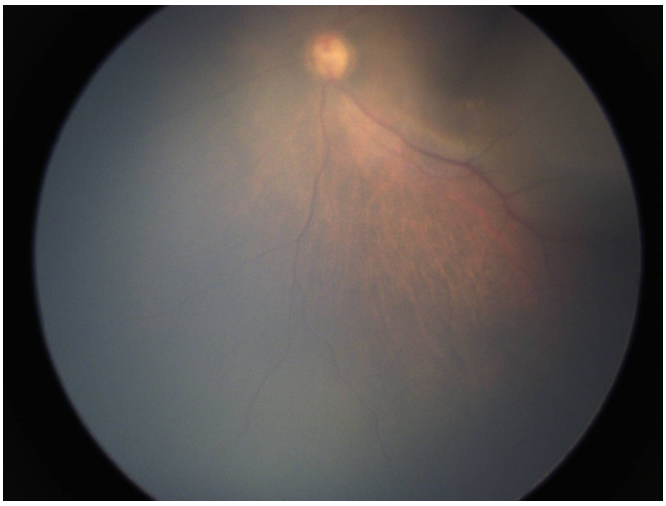
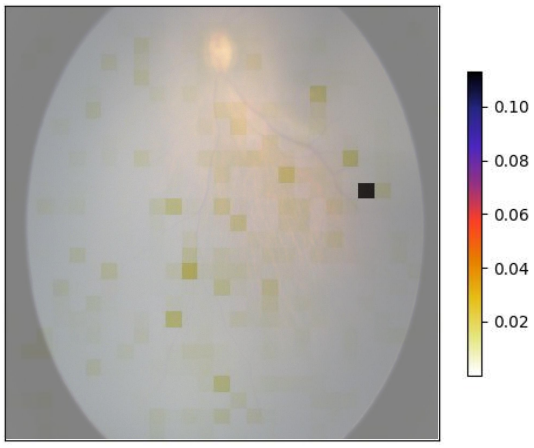


Figure 11. Instance weights of a sample with the label of WS.



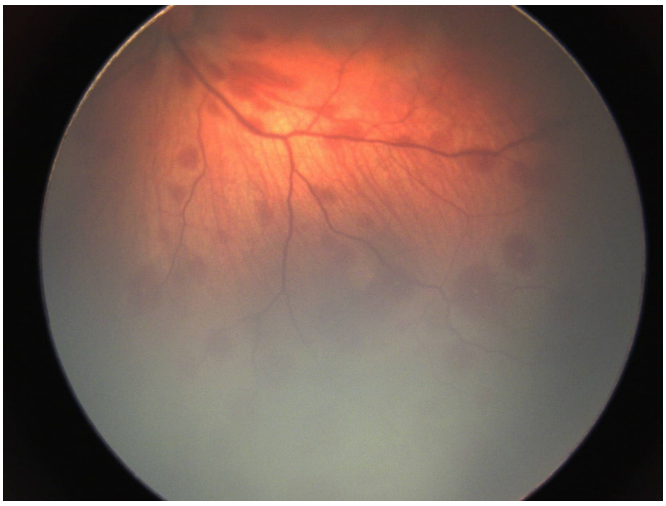
(a) An instance with no obvious abnormality.



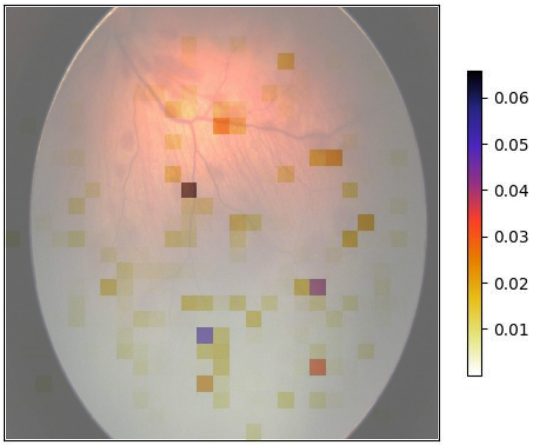
未见明显异常。

(b) Attention map overlay with Chinese description.

Figure 12. Most weighted image and attention distribution of aggregator. (a) Raw fundus image input. (b) Attention map overlay demonstrating the model's region-specific focus, where deeper hues correspond to higher attribution scores.



(a) An instance of RH.

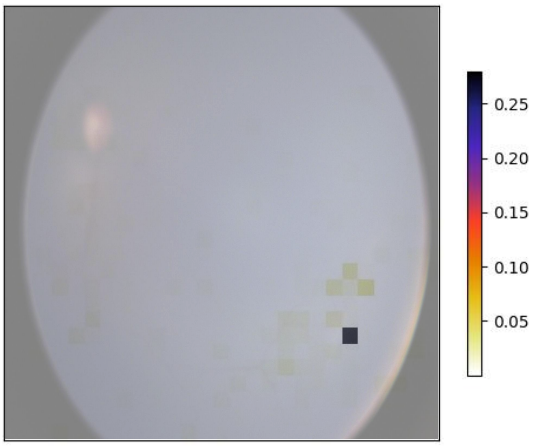


颞下方网膜可见沿视网膜神经纤维走向排列的浅层视网膜出血灶，呈红色。

(b) Attention map overlay with Chinese description.

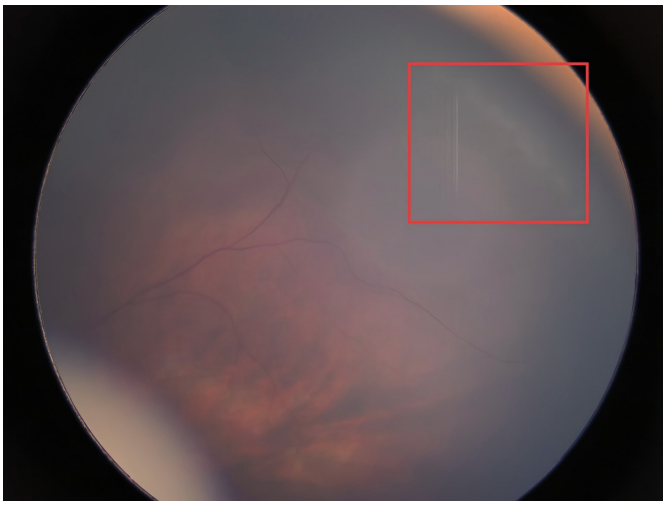
Figure 13. Most weighted image and attention distribution of aggregator. (a) Raw fundus image input. (b) Attention map overlay demonstrating the model's region-specific focus, where deeper hues correspond to higher attribution scores.

(a) An instance of ROP.

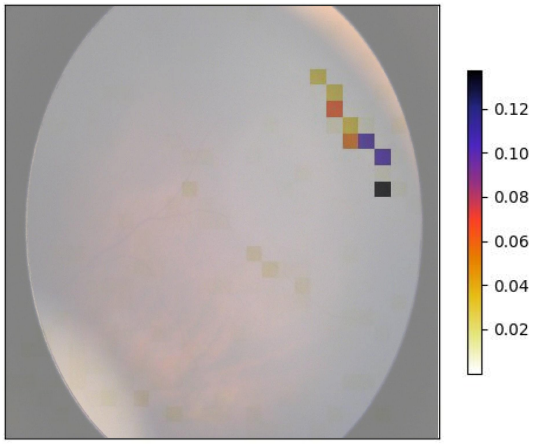


(b) Attention map overlay with Chinese description.

Figure 14. Most weighted image and attention distribution of aggregator. (a) Raw fundus image input. (b) Attention map overlay demonstrating the model's region-specific focus, where deeper hues correspond to higher attribution scores.



(a) An instance of WS.



(b) Attention map overlay with Chinese description.

Figure 15. Most weighted image and attention distribution of aggregator. (a) Raw fundus image input. (b) Attention map overlay demonstrating the model's region-specific focus, where deeper hues correspond to higher attribution scores.