Similarity-as-Evidence: Calibrating Overconfident VLMs for Interpretable and Label-Efficient Medical Active Learning

Supplementary Material

This document provides additional details and results to support the claims made in the main paper. Unless otherwise specified, all experiments follow the same protocol and hyperparameters described in the main text. We organize the supplementary material as follows:

- **Sec. 6** details the network architecture of the Similarity Evidence Head (SEH).
- **Sec. 7** presents comprehensive ablation studies, including the SEH loss design, sensitivity analyses of hyperparameters β and ϵ, the effect of context length M, and a validation of the acquisition schedule.
- **Sec. 8** demonstrates the visual interpretability of SaE through qualitative Grad-CAM comparisons and uncertainty maps.
- **Sec. 9** describes the pipeline used to collect and curate PubMed-augmented prompts and illustrates it with examples on BTMRI.

6. Similarity Evidence Head Architecture

Fig. 6 illustrates the SEH, a lightweight dual-branch network designed to map frozen VLM outputs to a scalar evidence strength λ. The architecture comprises an image branch that processes high-dimensional embeddings $\mathbf{x}$ through two stacked blocks to extract deep semantic cues and a similarity branch that encodes the raw similarity vector $\mathbf{s}$ via a single block. These feature streams are concatenated and fused by a final linear projection followed by a Softplus activation, strictly enforcing the positivity constraint ($\lambda > 0$) required for Dirichlet parameterization.

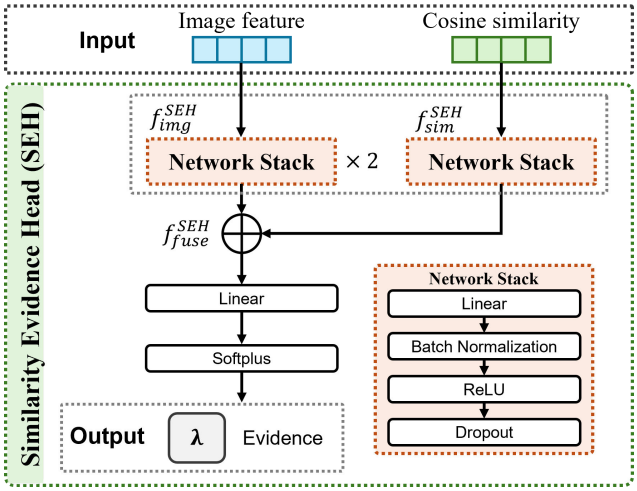

Figure 6. Architecture of the Similarity Evidence Head (SEH). SEH employs a dual-branch design to process image features $\mathbf{x}$ and similarity scores $\mathbf{s}$. The image branch consists of two stacked MLP blocks (Linear-BN-ReLU-Dropout) to extract deep semantic cues, while the similarity branch uses a single MLP block. These features are concatenated and fused via a final linear layer followed by a Softplus activation to ensure the output evidence strength λ is strictly positive.

7. Ablation Analysis

7.1. Ablation on SEH Loss Design

We validate the design of our Similarity Evidence Head (SEH) loss function by decomposing it into two distinct aspects: the information sources (loss components) and the mathematical formulation (regression targets). All experiments are conducted on the BTMRI dataset [49] at a 20% budget.

Impact of Loss Components. The SEH is designed to act as a calibration bridge, connecting empirical classification difficulty (l_{cls}) with the VLM's intrinsic uncertainty ($H[\mathbf{p}]$). To isolate the contribution of each signal, we explicitly define the two components of the total SEH loss:

$$\mathcal{L}_{\text{SEH}} = \underbrace{\text{MSE}\left(\frac{1}{\lambda+\epsilon}, l_{\text{cls}}\right)}_{\mathcal{L}_{\text{diff}}} + \beta \underbrace{\text{MSE}\left(\lambda, \frac{1}{H[\mathbf{p}]+\epsilon}\right)}_{\mathcal{L}_{\text{ent}}}. \tag{9}$$

Here, $\mathcal{L}_{\text{diff}}$ forces the evidence strength to reflect ground-truth difficulty, while $\mathcal{L}_{\text{ent}}$ ensures consistency with the VLM's prior knowledge. We verify the necessity of this dual-objective design by training SaE with each component individually. Table 5 summarizes the results. Using only entropy ($\mathcal{L}_{\text{ent}}$) yields the worst performance among the variants we test (92.15%), suggesting that merely mimicking the VLM's existing confidence fails to correct its inherent overconfidence. Using only difficulty ($\mathcal{L}_{\text{diff}}$) improves accuracy to 92.90% but lacks the distributional prior from the pre-trained VLM. The full dual objective ($\mathcal{L}_{\text{SEH}}$) achieves the highest accuracy of 93.46%. This confirms that fusing empirical errors with semantic priors is crucial for robust evidence estimation.

Table 5. Ablation of SEH loss components on BTMRI [49]. $\mathcal{L}_{\text{diff}}$ is the difficulty matching term, and $\mathcal{L}_{\text{ent}}$ is the entropy consistency term. The combination yields the best performance.

Method	Loss Formulation	Acc (%)	NLL
Entropy Only	$\mathcal{L}_{\text{SEH}} = \mathcal{L}_{\text{ent}}$	92.15	0.468
Difficulty Only	$\mathcal{L}_{\text{SEH}} = \mathcal{L}_{\text{diff}}$	92.90	0.441
SaE (Dual)	$\mathcal{L}_{\text{SEH}} = \mathcal{L}_{\text{diff}} + \beta\mathcal{L}_{\text{ent}}$	**93.46**	**0.425**

Impact of Regression Forms. We next investigate the mathematical form of the regression target. Our proposed method uses an inverse form (λ^{-1}), which naturally maps high uncertainty to low evidence values close to zero. We compare this against a standard logarithmic form ($-\log\lambda$), often used in uncertainty quantification to regress variance parameters. Table 6 compares these two variants. The log form achieves a competitive accuracy of 93.35%. However, our proposed inverse form slightly outperforms it (93.46%) and provides better calibration (lower NLL). We hypothesize that the inverse form provides stronger gradient signals for hard samples where λ approaches zero. We therefore adopt the inverse form as our default design.

Table 6. Comparison of regression target forms on BTMRI [49]. The inverse form (regressing $1/\lambda$) slightly outperforms the log form (regressing $-\log\lambda$) in both accuracy and calibration.

Regression Variant	Target Transform	Acc (%)	NLL
Log-form	$-\log(\lambda)$	93.35	0.432
Inverse-form (Ours)	$(\lambda+\epsilon)^{-1}$	**93.46**	**0.425**

7.2. Effect of β

The hyperparameter β in Eq. 3 controls the trade-off between fitting the empirical classification difficulty and aligning with the VLM's intrinsic entropy. We vary β in $\{0.1, 0.3, 0.5, 0.7, 1.0\}$ while fixing $\epsilon = 10^{-3}$, and measure Top-1 accuracy, NLL, and ECE at a 20% budget on BTMRI [49] and RETINA [39, 53]. Results are shown in Table 7. We observe that performance is stable across a broad range of β. In practice, we set $\beta = 0.5$ for all datasets to strike a balance between the two learning objectives.

Table 7. Sensitivity of SaE to the loss weight β. Performance is robust across a wide range of β, with the optimal trade-off consistently observed at $\beta = 0.5$. Extreme values (0.1 or 1.0) tend to degrade both calibration (NLL/ECE) and accuracy.

Dataset	Metric	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 1.0$
BTMRI	Acc (%)	92.82	93.18	**93.46**	93.24	92.95
	NLL	0.452	0.435	**0.425**	0.431	0.448
	ECE	0.029	0.025	**0.021**	0.024	0.027
RETINA	Acc (%)	73.54	74.65	**75.22**	74.92	74.18
	NLL	0.535	0.508	**0.492**	0.501	0.519
	ECE	0.047	0.041	**0.039**	0.040	0.044

7.3. Effect of ϵ

We next fix $\beta = 0.5$ and evaluate the impact of the numerical stability constant ϵ in the SEH loss on the BTMRI dataset. We vary $\epsilon \in \{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$ and report the Top-1 accuracy at a 20% budget. As shown in Table 8, SaE maintains stable performance for ϵ between 10^{-4} and 10^{-2}. We therefore fix $\epsilon = 10^{-3}$ in all experiments to ensure numerical stability without compromising accuracy.

Table 8. Sensitivity of SaE to the numerical stability parameter ϵ on BTMRI. SaE is largely insensitive to ϵ within a reasonable range (10^{-4} to 10^{-2}), with peak performance at the default setting.

ϵ	1×10^{-4}	5×10^{-4}	1×10^{-3}	5×10^{-3}	1×10^{-2}
Acc (%)	93.41	93.44	**93.46**	93.38	93.25

7.4. Effect of Context Length

The length of the learnable context vectors M acts as a critical hyperparameter governing the trade-off between adaptation capacity and overfitting risk. We investigate the sensitivity of SaE to this parameter by varying $M \in \{4, 16, 32, 64\}$ on DermaMNIST [15, 66] and BTMRI [49] at a fixed budget ($\rho = 0.2$). Table 9 summarizes the results. We observe that setting $M = 16$ consistently yields the optimal performance across both datasets. Shorter contexts ($M = 4$) appear insufficient to capture the necessary domain-specific semantics for fine-grained medical classification. Conversely, increasing the context length to 32 or 64 leads to a noticeable performance degradation. This drop likely stems from the model overfitting to the small labeled set available in active learning. Consequently, we adopt $M = 16$ as the default setting to ensure robust few-shot adaptation without overfitting.

Table 9. Impact of context length M on AL performance ($\rho = 0.2$). A moderate length ($M = 16$) achieves the best accuracy by balancing semantic capacity and overfitting. Performance drops at larger lengths ($M = 32, 64$) due to overfitting on limited active learning data.

Dataset	$M = 4$	$M = 16$	$M = 32$	$M = 64$
DermaMNIST	78.86	**80.21**	79.57	77.26
BTMRI	92.22	**93.46**	92.93	91.56

7.5. Effect of Acquisition Schedule

The dynamic interaction between vacuity and dissonance is critical for maximizing AL efficiency across different phases. We validate our proposed schedule by comparing it against three static baselines on the BTMRI dataset [49]. Specifically, we evaluate: (i) a vacuity-only strategy where

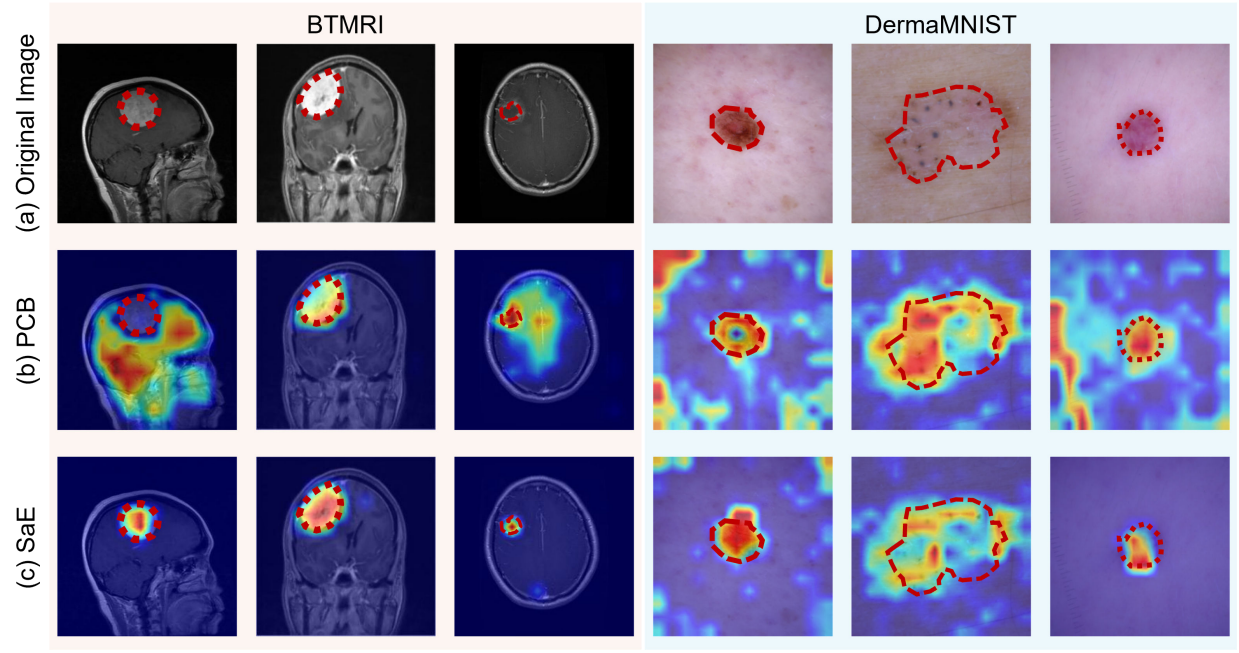

Figure 7. Visual interpretability comparison on BTMRI [49] and DermaMNIST [15, 66]. We visualize Grad-CAM [59] activation maps for the PCB [4] and our SaE. The red dashed contours indicate the ground-truth lesion regions. (a) Original input images. (b) PCB attention is often scattered, focusing on irrelevant background regions or failing to cover the entire lesion. (c) SaE generates highly focused and accurate attention maps that align closely with the pathological regions, confirming that our evidence-calibrated strategy successfully localizes clinical features.

$w_v(t) = 1$ and $w_d(t) = 0$; (ii) a dissonance-only strategy where $w_v(t) = 0$ and $w_d(t) = 1$; and (iii) a static balanced strategy where $w_v(t) = w_d(t) = 0.5$ throughout all rounds. Table 10 summarizes the performance differences. The dissonance-only strategy yields the poorest results (89.12%). This significant drop confirms that prioritizing ambiguous boundary cases before the model establishes a solid knowledge base leads to severe cold-start failure. The Vacuity-only strategy performs competitively in early stages but eventually plateaus (92.55%). This suggests that relying solely on exploration fails to refine decision boundaries in later rounds. The static balanced approach improves over single-factor methods but still lags behind our dynamic schedule. These results strongly support our explore-then-refine hypothesis, where the model benefits from prioritizing high-vacuity samples early and high-dissonance samples later.

Table 10. Ablation of acquisition schedules on BTMRI [49]. The dynamic schedule outperforms all static variants. Dissonance-only fails due to cold-start instability, while Vacuity-only lacks late-stage refinement. Our dynamic strategy optimally bridges these two needs.

Schedule Strategy	$w_v(t)$	$w_d(t)$	**Acc (%)**	**NLL**
Dissonance-only	$= 0$	$= 1$	89.12	0.584
Vacuity-only	$= 1$	$= 0$	92.55	0.460
Static balanced	$= 0.5$	$= 0.5$	92.85	0.445
SaE (Dynamic)	$\mathbf{1 - \frac{t-1}{T-1}}$	$\mathbf{\frac{t-1}{T-1}}$	**93.46**	**0.425**

8. Visual Interpretability

To qualitatively validate the source of our model's performance and calibration, we employ Grad-CAM [59] to visualize the regions that contribute most to the predictions. Fig. 7 compares the activation maps of our SaE framework against the PCB [4] on two representative datasets: BTMRI [49] (brain tumors) and DermaMNIST [15, 66] (skin lesions). As illustrated in the second row of Fig. 7, the PCB often exhibits scattered attention. It frequently focuses on irrelevant background structures (e.g., the skull boundary in MRI or healthy skin texture) rather than the pathological lesion. This behavior suggests that its high confidence scores may stem from spurious correlations rather than true clinical features, which helps explain its tendency toward overconfidence on out-of-distribution samples. In contrast, the third row shows that SaE generates highly focused at-

tention maps that align closely with the ground-truth lesion contours (indicated by red dashed lines). Whether for the intricate boundaries of a glioma or the pigmented structure of a melanoma, SaE consistently attends to the clinically relevant regions. This visual evidence supports our claim that the evidential calibration mechanism guides the model to leverage correct semantic features, thereby providing a transparent and trustworthy basis for its uncertainty estimates.

9. PubMed-Augmented Prompts Collection

To bridge the semantic gap between generic VLMs and specialized medical terminology, we construct class-specific imaging prompts by augmenting each category name with descriptions mined from PubMed [1]. PubMed serves as the primary source of domain knowledge, while a large language model is used only to summarize and rewrite the retrieved texts into concise prompts. For each class c_k in the dataset, we perform the following steps:

1. Query PubMed [1] for the top 20-50 articles that mention c_k together with imaging-related terms (e.g., "MRI", "X-ray", "lesion").
2. Export the retrieved records (titles and abstracts) and extract sentences that contain both the class name c_k and at least one imaging keyword (e.g., "mass", "enhancement", "hyperintense"). The selected sentences are concatenated into a short context document D_k for each class.
3. Use Google Gemini 2.5 Pro[1] to process D_k together with a structured instruction prompt that asks for 3-5 sentences describing the morphology, signal characteristics, and anatomical location of c_k based on the provided PubMed context.
4. Deduplicate the resulting candidate sentences using cosine similarity (threshold > 0.9) and discard candidates that are overly long (more than 30 words) or contain non-standard terminology.
5. Select approximately 10 diverse descriptions per class from the remaining pool, and have a board-certified radiologist validate the final set.

Extraction Prompt Template. The following template is used when querying Gemini for each class name with its PubMed-derived context:

```
"Extract 3-5 imaging feature descriptions
for [class_name]. Focus on: (1) morphology
(shape, borders), (2) signal patterns
(T1/T2, enhancement), (3) location and
tissue effects."
```

[1] https://deepmind.google/technologies/gemini/

9.1. Example Prompts: BTMRI Dataset

Below we show the resulting prompts for the four classes in the BTMRI brain tumor MRI dataset, including the normal brain category.

Glioma Tumor ($\delta_k = 10$):

```
"In the image, a glioma tumor often
appears as an irregularly shaped mass
with indistinct borders blending into the
surrounding brain tissue."

"The photo shows a lesion with
heterogeneous signal intensity, indicating
a mixture of different tissue types within
the glioma."

"A key feature in the image is a
ring-like enhancement pattern after the
administration of contrast agent, often
seen in high-grade gliomas."

"This image displays significant swelling,
known as vasogenic edema, in the brain
tissue surrounding the glioma tumor."

"A glioma tumor is visible as a mass
causing a 'mass effect', meaning it
displaces or deforms adjacent normal brain
structures."

"The tumor in the photo has infiltrated
the white matter tracts, which is a
characteristic feature of a glioma."

"This is an image of an intra-axial lesion,
meaning the glioma tumor originates from
within the brain parenchyma itself."

"The photo depicts a mass with central
necrosis or cystic components, appearing
as a dark area within the tumor."

"On this scan, the glioma appears as a
poorly-defined, infiltrative lesion within
the cerebral hemisphere."

"The image shows a brain with a glioma,
characterized by its invasive nature and
lack of a clear capsule."
```

Meningioma Tumor ($\delta_k = 10$):

```
"The image displays a meningioma as
a well-defined, distinctly bordered
mass located outside the brain tissue
(extra-axial)."

"A characteristic feature in the photo is
the tumor's broad attachment to the dura,
the outer lining of the brain."

"This is a photo of a meningioma showing
intense and uniform enhancement after
contrast injection, appearing as a brightly
lit mass."

"In the image, a 'dural tail sign' is
```

visible, which looks like a tail of enhancement extending from the tumor along the dura."

"The photo shows a meningioma causing compression and displacement of the adjacent brain without invading it."

"A meningioma often appears as a rounded or lobulated mass located on the surface of the brain."

"The tumor in this image is isointense to gray matter on T1-weighted images, meaning it has a similar shade before contrast."

"This image shows a meningioma that may contain calcifications, which appear as very dark or bright spots depending on the imaging sequence."

"The photo depicts a smoothly marginated mass consistent with a meningioma, pressing on the cerebral cortex."

"Visible in the scan is a dural-based mass, a classic presentation of a meningioma tumor."

Pituitary Tumor ($\delta_k = 10$):

"The image shows a well-defined, rounded mass located within the sella turcica, the bony cavity at the base of the skull where the pituitary gland sits."

"This photo depicts an expansion of the sella turcica caused by a pituitary tumor."

"A pituitary adenoma is visible in the scan, often appearing as a lesion with a signal intensity different from the normal pituitary gland."

"In the image, the pituitary tumor can be seen extending upwards (suprasellar extension), potentially compressing the optic chiasm."

"This is a photo of a pituitary macroadenoma, defined as a tumor larger than 10mm in diameter, filling the pituitary fossa."

"The image shows a pituitary tumor that enhances less avidly and more slowly than the surrounding normal pituitary gland after contrast."

"On this sagittal view, a mass is clearly visible within the pituitary fossa, characteristic of a pituitary tumor."

"The photo shows a lesion at the base of the brain consistent with a pituitary tumor, which can sometimes be cystic or hemorrhagic."

"This scan reveals a pituitary tumor causing thinning and remodeling of the bone of the sella turcica."

"The image displays a distinct mass in the sellar region, which is the typical location for a pituitary tumor."

Normal Brain ($\delta_k = 10$):

"The image displays a normal brain with clear differentiation between the gray matter on the outside and the white matter on the inside."

"In this photo, the brain structures are symmetrical on both the left and right sides, with no evidence of displacement."

"A normal brain scan shows the ventricles, the fluid-filled spaces, as well-defined structures of normal size and shape."

"The image shows no signs of abnormal masses, lesions, or growths within the brain tissue."

"This is a photo of a healthy brain where there is no abnormal enhancement after the administration of a contrast agent."

"In the image, the sulci and gyri, which are the grooves and folds of the brain cortex, appear normal and are not effaced."

"The scan shows a brain with no evidence of swelling (edema), hemorrhage, or fluid collections."

"This photo displays normal brain anatomy with all major structures, like the cerebrum, cerebellum, and brainstem, appearing unremarkable."

"A normal brain image is characterized by the absence of any pathological findings such as tumors, infarcts, or inflammation."

"The image shows a brain with normal signal intensity throughout, without any bright or dark spots that would suggest pathology."