# Assessing Large Vision Models with GPT-5 for Advancing Zero-Shot Capabilities

Jun Yeong Lee, Gyeong Hee Jung, and Dong Seog Han
*Department of Electronic and Electrical Engineering*
*Kyungpook National University, Daegu, Republic of Korea*
{ddr37105, ghjung, dshan}@knu.ac.kr

***Abstract*—While in-domain fine-tuning remains one approach to adapt models under data scarcity, the primary goal of foundation models is to reduce unnecessary training by leveraging prompting, as demonstrated in large language models (LLMs). However, while text-to-text or vision-to-text tasks have been extensively studied with well-established evaluation metrics, vision-to-vision tasks remain comparatively underexplored, and reliable evaluation methodologies are still lacking. To address this gap, we propose an evaluation-driven framework that utilizes GPT-5 to analyze the predictions of vision-only models. Based on these evaluations, we employ a vision foundation model to perform historic map retrieval, where current satellite images are retrieved by embedding both historic and modern images and matching them through cosine similarity without additional training. As a preliminary step, we first evaluate an image-to-text task, circuit diagram classification, where GPT-based evaluation guides whether performance can be improved purely through prompting without fine-tuning. Building on these findings, we propose a tailored training-free adaptation method for the image-to-image setting of historic map retrieval. Our approach achieves a Recall at 1 score of 30.3% and demonstrates the practicality of evaluation-guided prompting for vision-only foundation model adaptation.**

***Index Terms*—large vision models, zero-shot learning, historic map retrieval, circuit diagram classification, evaluation-guided prompting**

## I. Introduction

In computer vision, the rapid progress of deep learning has largely centered on supervised models trained on large-scale labeled datasets. These models tend to show strong performance only on the data they were trained on, and recent attention has therefore shifted to whether they can be applied to previously unseen patterns, such as inspecting newly developed products. Recent foundation models, often referred to as large vision models (LVMs), are trained on massive and heterogeneous datasets, enabling them to capture broad semantic knowledge. By leveraging such large-scale and diverse supervision, these models exhibit robust zero-shot capabilities across a variety of visual tasks, ranging from classification and detection to segmentation and retrieval, without the need for task-specific fine-tuning [1]–[3].

In real-world implementations, LVMs are sometimes trained with additional labeled datasets to improve performance in specialized domains. However, as shown in works such as DINO [4], directly retraining all weights for domain adaptation can disrupt previously learned representations and degrade generalization. To mitigate this, recent approaches have explored partial tuning strategies, where only a subset of parameters is updated while keeping most of the backbone frozen [5]–[8]. In transformer architectures, domain adaptation is often performed by training the classification head to learn new class embeddings or by decomposing weight matrices into smaller trainable components for fine-grained tuning. While these approaches reduce computational cost and mitigate catastrophic forgetting [9], they inherently replace the original embedding-matching mechanism of zero-shot inference with a supervised classification structure. This shift highlights a fundamental trade-off between efficiency, specialization, and the preservation of the embedding-based inference principle.

In a related field, large language models (LLMs) address the challenges of retraining and fine-tuning through prompt-based adaptation. Instead of updating model weights—which in LVMs often introduces risks such as computational overhead or degradation of pre-trained representations. In practice, personalization is achieved by augmenting prompts, often with retrieval-augmented generation (RAG), allowing users to adapt models to specific tasks without additional training. Moreover, prompting serves not only as an efficient method for task adaptation but also as a scalable and flexible framework for evaluation [10]–[12]. The effectiveness of this paradigm relies on the quality of text tokens, as stronger prompts derived from fixed sentences yield more reliable outputs [13]. Motivated by this perspective, this paper investigates the use of GPT-driven prompting to evaluate pure vision models, aiming to improve performance without retraining. The approach can be viewed as transferring vision-to-text outputs from LLMs into vision-to-vision prompts, thereby establishing a cross-modal connection without the need for additional training.

Our objective is to evaluate the applicability of pure vision models in challenging tasks such as circuit diagram classification and historic map retrieval [14], both of which require aligning heterogeneous visual patterns. Unlike language-centric domains where prompting is naturally applicable, pure vision tasks do not readily allow the use of GPT-style prompting. In this paper, we novelly introduce GPT-driven evaluation as a means to analyze misclassifications and, based on these insights, propose how performance can be improved in pure vision models without additional training.

This work was supported by IITP–ITRC grant funded by MSIT (IITP-2025-RS-2020-II201808).

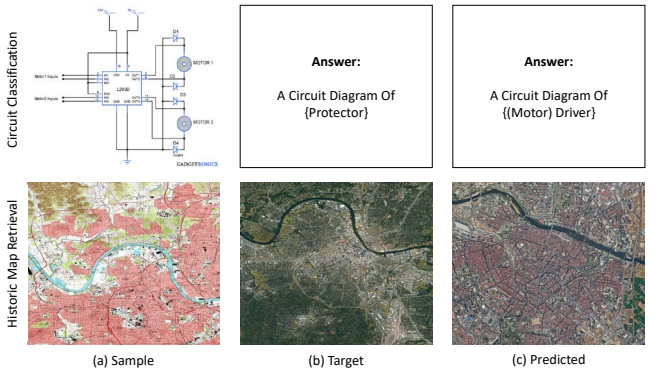


Fig. 1. Illustration of zero-shot retrieval outcomes using ImageBind. Subfigure (a) shows the input image to be classified, (b) indicates the correct target class, and (c) is the model's incorrect prediction.

The main contributions of this paper are summarized as follows:

- We present an evaluation protocol for LVMs using state-of-the-art LLMs as reference and introduce a vision-to-vision prompting approach to improve zero-shot classification without additional training.
- We validate the proposed approach through GPT-driven evaluation on the Low-Resource Large Vision Model Challenge, confirming its effectiveness in practical vision tasks.
- We contribute to improving the usability of Large Vision Models by demonstrating how GPT-driven evaluation can guide performance enhancement without retraining.

## II. Text-Based Evaluation

When applying large vision models (LVMs) to domain-specific tasks, evaluation is often limited to checking whether the output matches the target, without revealing the reasons for errors. As illustrated in Fig. 1, misclassified cases only indicate that the prediction is wrong, not why it failed. While domain experts may infer causes, such as recognizing component roles in a circuit, this is not feasible in automated evaluation. These limitations motivate a GPT-driven evaluation framework that goes beyond visual similarity and provides interpretable reasoning for misclassifications.

Recent advances such as ChatGPT have shown that large language models can perform multimodal tasks, including image analysis and online search. These capabilities build on progress in text–image embedding, where visual and linguistic features are aligned in a shared latent space for flexible recognition. A prominent example is CLIP [15], which jointly trains image and text encoders to maximize similarity for matched pairs and minimize it for mismatches. In zero-shot settings, categories are represented by templated prompts (e.g., "a photo of a {class}"), and predictions rely on cosine similarity in the embedding space. To further improve alignment, context optimization (CoOp) [16] replaces static prompts with learnable continuous context vectors. While these approaches enhance adaptability, they struggle to separate semantically

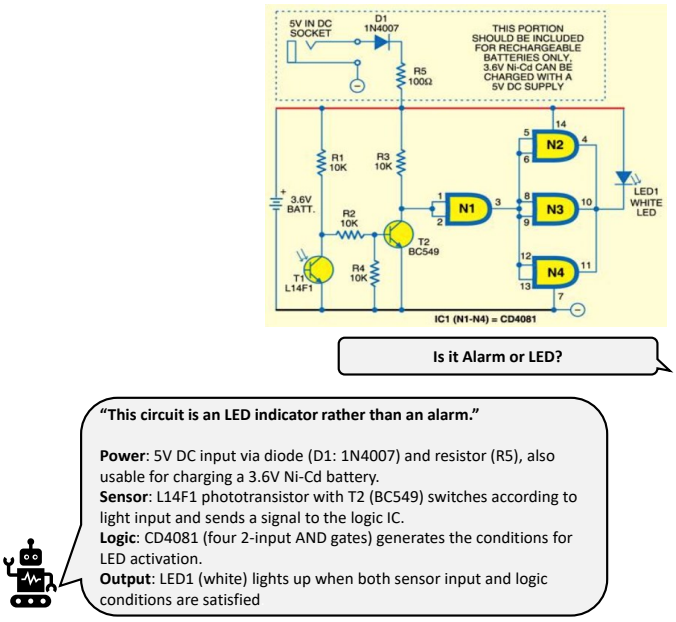


Fig. 2. Example of a misclassified sample where both the LVM encoder and the GPT-5 evaluation produced the same incorrect prediction. The circuit is not a simple "LED indicator" but an alarm circuit in which the L14F1 phototransistor detects incoming light, the CD4081 AND gate controls activation, and the white LED (LED1) serves as a visual alarm.

similar class names, highlighting the need to refine not only model parameters but also the structure and specificity of prompts used for inference.

In low-resource domains, such as those in the low-resource image transfer evaluation (LITE) benchmark [17], the limitations of zero-shot models are more evident. As shown in Fig. 2, both the LVM encoder and GPT evaluation can yield the same incorrect prediction. Unlike the LVM-only setting, however, GPT provides textual reasoning that explains the misclassification. For instance, it may label a circuit as an "LED indicator" because an LED is visible, yet fail to infer that the overall design functions as an alarm. With additional prompts, GPT can clarify that the L14F1 phototransistor detects incoming light, the CD4081 AND gate ensures conditional activation, and the white LED (LED1) serves as a visual alarm. This demonstrates that GPT-driven feedback not only identifies the cause of errors but also guides model improvement under limited supervision.

## III. Application on Vision-Only Dataset

In vision-only tasks, it is often necessary to reason over patterns that were learned in different styles. A representative example is historic map retrieval, where each map is drawn with a distinct artistic style, yet must be matched to its corresponding modern satellite image. Unlike text-driven settings, this is a pure vision-to-vision problem, which calls for new approaches beyond conventional embedding similarity. To better understand visual decision dynamics in such tasks, we evaluated retrieval behavior using GPT as a diagnostic proxy. As shown in Fig. 3, GPT exhibited classification

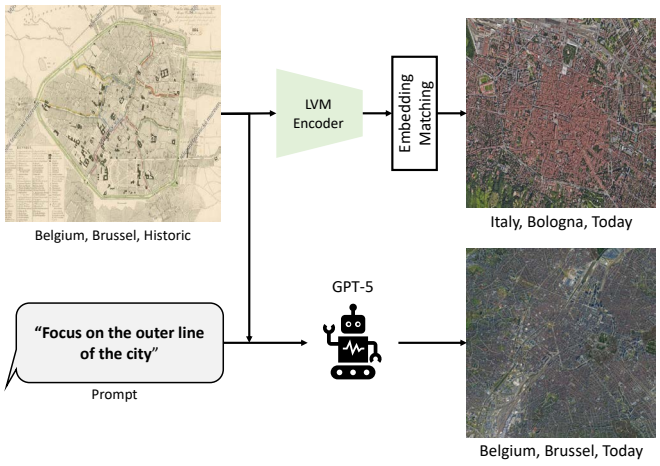


Fig. 3. Illustration of GPT-guided zero-shot retrieval in the historic map to satellite image matching task. The figure shows how GPT initially produces an incorrect prediction, but when provided with additional prompt information, it can reason about contextual cues and correctly identify the location as Belgium, Brussels.

patterns similar to those of LVM encoders, misidentifying the target satellite image due to an overreliance on radial grid patterns that dominated the embedding representation. However, when the prompt was modified to deprioritize radial layout and emphasize peripheral attributes, GPT's diagnostic evaluation showed that the correct image could be retrieved by focusing on outer road structures. This result indicates that GPT can reveal how global structural features—particularly dominant radial configurations—disproportionately influence cosine similarity outcomes, thereby underscoring the need to disentangle global patterns from finer contextual cues when designing vision-only similarity measures.

Motivated by the GPT evaluation results and [18], which revealed that similarity judgments were dominated by global radial structures while neglecting finer details, we adopt a pure-vision classification approach that explicitly disentangles global and local components. Historic maps are matched to modern satellite imagery through cosine similarity of visual features, while ImageBind embeddings are decomposed into complementary representations: global patterns are extracted via low-pass filtering and statistical averaging, and local variations are captured through high-pass residuals and frequency-domain separation using the fast Fourier transform (FFT). Formally, given embeddings $\mathbf{f}_i$ and $\mathbf{f}_j$ from two images, their similarity and decomposition are defined as:

$$\mathbf{f} = \mathbf{f}_{\text{low}} + \mathbf{f}_{\text{high}} \tag{1}$$

$$\text{sim}(\mathbf{f}_i, \mathbf{f}_j) = \frac{\langle \mathbf{f}_i, \mathbf{f}_j \rangle}{\|\mathbf{f}_i\|_2 \, \|\mathbf{f}_j\|_2} \tag{2}$$

where $\mathbf{f}$ denotes the original embedding, $\mathbf{f}^{\text{low}}$ and $\mathbf{f}^{\text{high}}$ represent the low- and high-frequency components obtained through decomposition, and $\text{sim}(\mathbf{f}_i, \mathbf{f}_j)$ is the cosine similarity between two embeddings $\mathbf{f}_i$ and $\mathbf{f}_j$.

As illustrated in Fig. 5, our framework separates high-frequency and low-frequency components from visual embeddings, computes cosine similarity for each, and integrates them through a weighted combination, $\alpha \cdot \text{sim}_{\text{global}} + \beta \cdot \text{sim}_{\text{detail}}$. This hybrid formulation enables flexible emphasis on either structural layout or fine-grained components depending on task-specific demands. To further interpret model behavior, we employed GPT to analyze visual attention patterns and identify retrieval errors. The evaluation revealed that baseline models tend to focus excessively on dominant global features, such as radial grids, while overlooking peripheral contextual structures. In contrast, our method highlights outer road networks—features that GPT identified as critical for correctly distinguishing spatially similar yet functionally distinct regions. These results demonstrate that multi-scale visual similarity, supported by GPT-based interpretation, provides a more robust and interpretable foundation for pure-vision classification tasks.

## IV. Experimental Results

### A. Dataset and Evaluation Index

We conduct experiments on the low-resource image transfer evaluation (LITE) benchmark introduced by Zhang et al. [17]. The LITE benchmark includes three low-resource vision tasks—circuit diagram classification, historic map retrieval, and mechanical drawing retrieval—with only a few hundred labeled samples per task. For example, circuit diagram classification comprises 154 train, 100 validation, and 1,078 test images; historic map retrieval includes 102 train, 140 validation, and 409 test samples; In this work, we evaluate models in a zero-shot setting using only the test splits without any training or fine-tuning.

For performance assessment, we use standard retrieval metrics. Circuit diagram classification is measured by Top-1 and Top-5 accuracy, while historic map retrieval is evaluated using Recall@1 (R@1), Recall@5 (R@5), and mean rank (MnR). Top-1 accuracy indicates the proportion of samples for which the correct class is ranked first, whereas Top-5 accuracy measures whether the correct class appears within the top five predictions. Similarly, higher values of Top-$k$ and R@$k$ correspond to better retrieval accuracy, while a lower MnR reflects improved ranking quality.

### B. Experimental Results

We evaluate our approach using the pre-trained ImageBind-huge encoder [19]. For circuit diagram classification, we refine the text embeddings by adopting domain-specific prompts such as "signal transmitter circuit with power amplifier, oscillator, and antenna output stage" versus "signal receiver circuit with low-noise amplifier, mixer, demodulator, and antenna input stage," instead of generic templates like "a circuit diagram of {amplifier}." For historic map retrieval, we decompose the embeddings into global and local components and compute cosine similarity for each before combining them into a hybrid

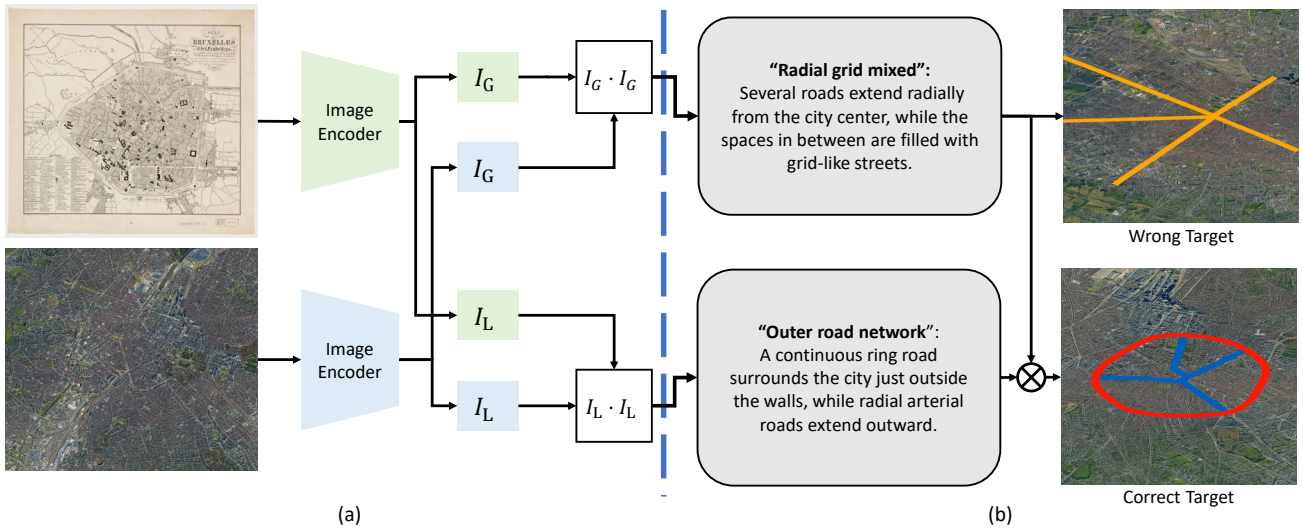


Fig. 4. Illustration of the proposed zero-shot enhancement. (a) shows the decomposition of image embeddings into high-frequency and low-frequency components, followed by separate cosine similarity computations. (b) presents the retrieval outcome, where the baseline model attends only to radial grid structures, whereas the proposed method captures peripheral road networks that are functionally more discriminative.

similarity score, enabling the model to capture both large-scale layouts and finer contextual cues.

The results are summarized in Table I. For circuit diagram classification, our domain-specific prompts improved Top-1/Top-5 accuracy from 19.3/45.1% to 21.2/50.2%. For historic map retrieval, our method raised R@1 from 28.1% to 30.3% and achieved a lower MnR (10.3 vs. 13.4) while maintaining comparable R@5 performance. These results confirm that global–local disentanglement and prompt refinement enhance zero-shot evaluation.

## V. Discussion and Conclusion

This paper highlights the increasing use of LVMs for domain-specific tasks without training. In such settings, the quality and design of prompts play a critical role in determining performance. However, relying solely on vision-to-vision similarity without textual prompts remains challenging, as current models still struggle to capture domain-specific semantics in a purely visual manner. This indicates that vision-only prompting strategies are not yet sufficient to achieve robust performance.

Nevertheless, as demonstrated in this paper, employing GPT-driven evaluation provides a valuable diagnostic tool that exposes the underlying causes of misclassification and guides the refinement of domain-specific prompts. This underscores the importance of systematic evaluation when deploying large vision models in specialized tasks. Furthermore, our findings confirm that even without additional fine-tuning, LVMs can deliver meaningful performance in low-resource, domain-specific settings through zero-shot classification supported by auxiliary evaluation mechanisms. In this sense, evaluation is not only a means of assessing performance but also a key enabler for making zero-shot LVMs practical in real-world applications.

## References


[1] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[3] X. Chen, K. Fang, Q. Yu, A. Zeng, Z. Xu, and B. Zhou, "Aim: Adapting image models for efficient multimodal learning," *arXiv preprint arXiv:2402.10354*, 2024.

[4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[6] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European conference on computer vision*. Springer, 2022, pp. 709–727.

[7] D. Zhang, T. Feng, L. Xue, Y. Wang, Y. Dong, and J. Tang, "Parameter-efficient fine-tuning for foundation models," *arXiv preprint arXiv:2501.13787*, 2025.

[8] Z. Fu, H. Yang, A. M.-C. So, W. Lam, L. Bing, and N. Collier, "On the effectiveness of parameter-efficient fine-tuning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 11, 2023, pp. 12 799–12 807.

[9] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] G. Team, R. Anil, S. Borgeaud, Y. Wu, J. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. Dai, A. Hauth *et al.*, "Gemini: A family of highly capable multimodal models, 2024," *arXiv preprint arXiv:2312.11805*, vol. 10, 2024.

[12] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language

TABLE I
IMAGEBIND PERFORMANCE ON CIRCUIT DIAGRAM CLASSIFICATION AND HISTORIC MAP RETRIEVAL.

| | **Circuit Diagram Classification** | | **Historic Map Retrieval** | | |
|---|---|---|---|---|---|
| **ImageBind [19]** | Top-1 ↑ | Top-5 ↑ | R@1 ↑ | R@5 ↑ | MnR ↓ |
| Zero-Shot Transfer [17] | 19.3 | 45.1 | 28.1 | 62.1 | 10.1 |
| + AdaptFormer [17] | 19.8 | 45.5 | 30.3 | 62.6 | 13.4 |
| Ours | 21.2 | 50.2 | 30.3 | 61.4 | 10.3 |

models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.

[13] J.-L. Peng, S. Cheng, E. Diau, Y.-Y. Shih, P.-H. Chen, Y.-T. Lin, and Y.-N. Chen, "A survey of useful llm evaluation," *arXiv preprint arXiv:2406.00936*, 2024.

[14] Y. Zhang, H. Doughty, and C. G. Snoek, "Low-resource vision challenges for foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 956–21 966.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[16] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Y. Zhang, H. Doughty, and C. G. M. Snoek, "Low-resource vision challenges for foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 21 956–21 966.

[18] R. Zeng, C. Han, Q. Wang, C. Wu, T. Geng, L. Huangg, Y. N. Wu, and D. Liu, "Visual fourier prompt tuning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 5552–5585, 2024.

[19] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 180–15 190.

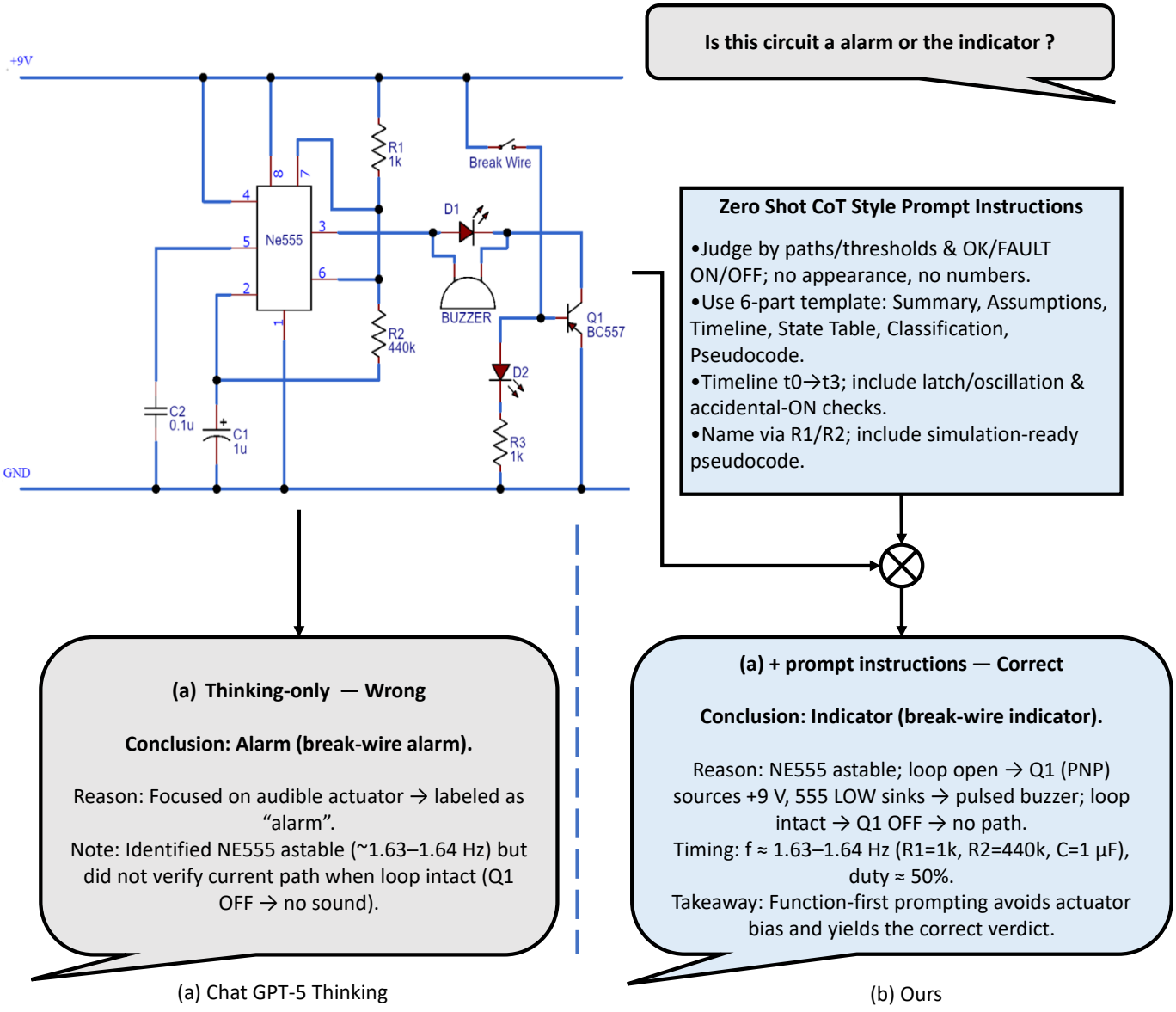


Fig. 5. Illustration of circuit reasoning outcomes. (a) ChatGPT-5 thinking-only approach misclassifies the design as an *alarm* due to actuator bias. (b) With zero-shot CoT style prompt instructions, the same circuit is correctly classified as a break-wire *indicator*, showing that function-first prompting yields the correct verdict.