

# DeepSymbol: A Multimodal Framework for Cultural Symbol Understanding

Sendong Liang

Auckland University of Technology  
Auckland, New Zealand  
sendong.liang@autuni.ac.nz

Parma Nand

Auckland University of Technology  
Auckland, New Zealand  
parma.nand@aut.ac.nz

Wei Qi Yan

Auckland University of Technology  
Auckland, New Zealand  
weiqi.yan@aut.ac.nz

**Abstract**—Cultural symbols carry important cultural meanings which can appear as static decorative patterns or dynamic gestures. Sometimes they can be difficult to recognise and interpret because they often require domain-specific knowledge. This paper asks how effectively static and dynamic cultural symbols can be recognised, and how their cultural meanings can be interpreted with appropriate evidence. To address these questions, we propose DeepSymbol, a unified framework that combines specialised visual perception with domain-specific semantic adaptation and evidence-aware reasoning. The framework uses two separate modelling strategies for static and dynamic symbols. Spatial modelling is used for static symbols, while temporal modelling is used for dynamic gestures. External cultural evidence is then used to support semantic interpretation. Experiments show strong performance across both recognition and semantic understanding tasks, with static recognition reaching 0.934 mAP@0.5 and dynamic gesture recognition reaching 0.889 accuracy. The complete framework achieves 0.893 semantic accuracy while reducing the hallucination rate from 10.9% to 4.8%. These results demonstrate the value of combining specialised perception and evidence-supported interpretation for structured cultural symbol understanding.

**Index Terms**—cultural symbol understanding, fine-grained visual recognition, Swin Transformer V2, LoRA, retrieval-augmented generation

## I. INTRODUCTION

Cultural symbols can appear as culturally meaningful visual or behavioural signs. Static decorative patterns depend mainly on spatial structure and subtle inter-class differences, while dynamic gestures depend on temporal ordering and action transitions. General-purpose models may struggle with specialised cultural content and symbolic meaning [1].

Recent advances in vision and multimodal learning provide strong foundations. Swin Transformer and Swin Transformer V2 improve spatial representation through hierarchical attention [2]; Transformer-based models capture temporal dependencies in video [3]; and multimodal large language models connect visual information with language [4]. However, specialised cultural recognition still involves fine-grained visual differences, temporal ambiguity, domain-specific terminology, long-tail knowledge, and the risk of fluent but weakly supported explanations. Parameter-efficient adaptation and retrieval provide complementary solutions [5]–[7].

This paper presents *DeepSymbol*, a unified framework for cultural symbol understanding. It integrates static and dynamic recognition with semantic interpretation: spatial modelling is

used for static patterns, while temporal modelling is used for dynamic gestures.

The main contributions are threefold:

- 1) A unified framework for static and dynamic cultural symbols, using spatial modelling for fine-grained patterns and temporal modelling for gesture sequences.
- 2) Task-specific semantic adaptation using LoRA to improve the interpretation of cultural terminology, visual-language mappings, and structured outputs.
- 3) Evidence-aware reasoning using dense retrieval, cross-encoder reranking, and evidence sufficiency gating to improve the reliability of cultural semantic interpretation.

## II. RELATED WORK

### A. Cultural Visual and Fine-Grained Recognition

CulturalVQA and VQArt-Bench show that multimodal models can vary in performance across cultural regions and symbolic visual tasks, highlighting the challenges of culturally grounded visual understanding [1], [8]. Fine-grained recognition is particularly relevant to cultural symbols because categories may share similar overall shapes while differing in local structures, composition, texture, and subtle appearance cues [9]. YOLO-based detectors provide practical object localisation [10], [11], while Swin Transformer and SwinV2 strengthen spatial representation through hierarchical window-based attention [2]. However, visual recognition primarily addresses localisation and category identification rather than cultural meaning. DeepSymbol therefore uses fine-grained visual recognition as the perception stage for subsequent cultural interpretation.

### B. Temporal and Multimodal Modelling

Dynamic gesture recognition requires information across frames because meaning may depend on movement order, duration, and transitions between postures. LSTM and GRU models capture temporal dependencies, while Transformer-based video models explicitly model relationships across spatial and temporal dimensions [3]. However, recognising a gesture sequence and interpreting its cultural meaning are different tasks. Multimodal language models connect visual information with language, but specialised cultural terminology and task-specific output formats may require adaptation [4]. DeepSymbol therefore separates temporal perception from

# Overall Research Framework of DeepSymbol

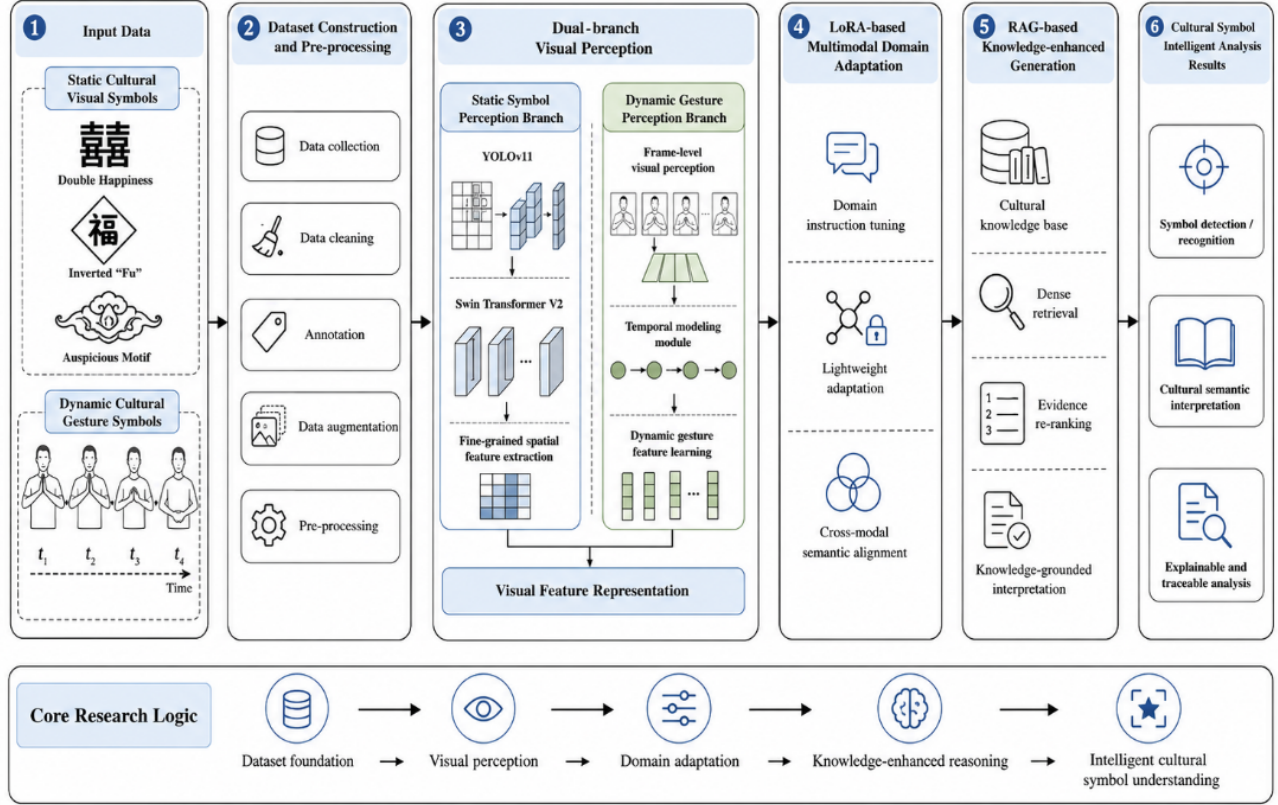


Fig. 1. Overall research framework and technical pipeline of DeepSymbol.

semantic interpretation, using temporal modelling for dynamic gestures and domain-specific adaptation for culturally grounded descriptions.

## C. Domain Adaptation and Retrieval-Augmented Reasoning

LoRA enables parameter-efficient adaptation by learning low-rank updates while keeping the principal model weights frozen, making it suitable for specialised domains with limited training data [5]. RAG introduces externally maintained knowledge into generation, while multimodal RAG extends this principle to image-text tasks [6], [7]. These approaches can provide domain-specific knowledge, but retrieving relevant information does not necessarily ensure that the evidence sufficiently supports the generated interpretation. DeepSymbol combines LoRA-based adaptation with dense retrieval and cross-encoder reranking, followed by an explicit evidence-sufficiency check before semantic generation.

## III. DEEPSYMBOL FRAMEWORK

### A. Overall Architecture

DeepSymbol consists of four stages: dataset construction, heterogeneous visual perception, domain-specific semantic adaptation, and knowledge-enhanced reasoning. Given an image or video input, the visual perception stage extracts structured information according to the modality, including cultural-symbol

categories, confidence scores, spatial locations, and temporal characteristics. The resulting visual information is then provided to the domain-specific semantic adaptation stage, where a multimodal language model is adapted to cultural terminology, visual-language mappings, and structured interpretation tasks using parameter-efficient fine-tuning. Finally, the knowledge-enhanced reasoning stage retrieves relevant external cultural evidence and incorporates it into the semantic generation process. This architecture separates visual perception, domain adaptation, and evidence grounding while maintaining a unified pipeline for static symbols and dynamic gestures.

### B. Heterogeneous Visual Perception

For static images, DeepSymbol uses a SwinV2-Tiny backbone within the YOLO11 pipeline for multi-scale feature extraction and detection. For dynamic gestures, the spatial representation is extended with ROI features, motion-aware scheduling, feature gating, and temporal self-attention. The process is summarised in Algorithm 1.

Here,  $D$  denotes detected instances,  $d_{\text{conf}}$  the detection confidence, and  $\tau_d$  the detection threshold. For dynamic input,  $B_t$ ,  $S_t$ , and  $F_t$  denote the detected regions, confidence scores, and ROI features at frame  $t$ , respectively. The static branch requires one backbone and detector pass, while the dynamic

---

**Algorithm 1: Heterogeneous Visual Perception**

---

**Input:** Static image or video clip  $I$   
**Output:** Visual representation  $R_v$   
**if**  $I$  is a static image **then**  
   $\{P_3, P_4, P_5\} \leftarrow \text{SwinV2Tiny}(I);$   
   $\mathcal{D} \leftarrow \text{YOLO11}_{det}(P_3, P_4, P_5);$   
   $\hat{Y}_s \leftarrow \emptyset;$   
  **foreach**  $d \in \mathcal{D}$  **do**  
    **if**  $d_{\text{conf}} \geq \tau_d$  **then**  
       $\hat{Y}_s \leftarrow \hat{Y}_s \cup \{d\};$   
   $R_v \leftarrow \hat{Y}_s;$   
**else**  
  **for**  $t = 1$  **to**  $T$  **do**  
     $\{P_3^t, P_4^t, P_5^t\} \leftarrow \text{SwinV2Tiny}(I_t);$   
     $(B_t, S_t) \leftarrow \text{YOLO11}_{det}(P_3^t, P_4^t, P_5^t);$   
     $F_t \leftarrow \text{ROIAlign}(P_4^t, B_t);$   
     $M \leftarrow \text{MotionMagnitude}(\{B_t\});$   
     $W \leftarrow \text{DynamicWindow}(M);$   
    **for**  $t = 1$  **to**  $T$  **do**  
       $G_t \leftarrow \text{Gate}(F_t, S_t);$   
       $F_t \leftarrow F_t \odot G_t;$   
   $R_v \leftarrow \text{TemporalTransformer}(\{F_t\}_{t=1}^T);$   
**return**  $R_v;$

---

branch scales with the number of sampled frames  $T$  and additionally incurs ROI, gating, and temporal-attention costs.

#### IV. DOMAIN-SPECIFIC SEMANTIC ADAPTATION

A multimodal instruction dataset of 6,400 samples was constructed, including 3,840 static-symbol samples and 2,560 dynamic-gesture samples. The train, validation, and test partitions contain 5,120, 640, and 640 samples, respectively, while retaining the partition membership of their underlying visual samples.

The semantic adaptation module uses DeepSeek-LLM-7B-Base as the language backbone. Visual representations are mapped into the language-model embedding space through a cross-modal projector, while the visual encoder and original language-model weights remain frozen. LoRA adapters are inserted primarily into the query and value projection layers. The evaluated LoRA ranks are 4, 8, and 16, with rank 8 selected as the principal configuration.

The adaptation objective is the autoregressive supervised fine-tuning loss

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{L} \sum_{t=1}^L \log p_{\Theta_L}(y_t \mid y_{<t}, I, X), \quad (1)$$

where  $L$  is the output sequence length,  $y_t$  is the target token,  $y_{<t}$  denotes the preceding target tokens, and  $p_{\Theta_L}$  is the conditional probability under the LoRA-adapted model.  $I$  and  $X$  denote the visual input and task instruction, respectively.

---

**Algorithm 2: Evidence-Grounded Cultural Interpretation**

---

**Input:** Structured visual-semantic output  $R$   
**Output:** Final interpretation  $Y$   
 $Q \leftarrow \text{BuildQuery}(R);$   
 $\mathcal{C} \leftarrow \text{DenseRetrieve}(Q, 20);$   
 $\mathcal{R} \leftarrow \text{CrossEncoderRank}(\mathcal{C});$   
 $E \leftarrow \text{TopK}(\mathcal{R}, 4);$   
**if**  $s_{\max}(E) \geq \tau_e$  **then**  
   $Y \leftarrow \text{Generate}(R, E);$   
**else**  
   $Y \leftarrow \text{GenerateLowerCertainty}(R, E);$   
**return**  $Y;$

---

The semantic interface uses structured fields for symbol type, class name, visual observation, basic semantics, and uncertainty, with an additional temporal summary for dynamic gestures. This structured output provides a consistent interface for subsequent evidence-grounded reasoning.

#### V. KNOWLEDGE-ENHANCED REASONING

The cultural knowledge base contains approximately 420 screened documents and 3,200 structured semantic chunks. Dense retrieval returns the top 20 candidates, which are reranked before the top four passages are provided to the generator. An evidence-sufficiency threshold  $\tau_e$  controls whether the retrieved evidence is sufficient for a confident interpretation: a lower threshold accepts more evidence but may increase weakly supported interpretations, whereas a higher threshold is more selective but may reject useful evidence. The threshold is fixed before final evaluation and is not tuned on the test set. The process is summarised in Algorithm 2. Here,  $Q$  is the retrieval query,  $\mathcal{C}$  the retrieved candidates,  $\mathcal{R}$  the reranked results,  $E$  the selected evidence,  $s_{\max}(E)$  the highest evidence score, and  $\tau_e$  the evidence-sufficiency threshold.

#### VI. EXPERIMENTAL SETUP

The study uses separate datasets for static symbols, dynamic gestures, semantic interpretation, and cultural knowledge, with their main statistics summarised in Table I. The static symbols cover eight categories: Double Happiness, Inverted Fu, Auspicious Cloud, Lotus, Peony, Dragon and Phoenix, Key Fret Thunder, and Chinese Knot. The dynamic set covers eight gestures: Fist-and-palm Salute, Rock-on, Finger-heart, Crossed fingers, ILY gesture, Call me, Air quotes, and Thumbs up. The semantic adaptation set contains 6,400 samples, split 80:10:10 into 5,120 training, 640 validation, and 640 test samples. Static and dynamic samples were manually annotated with category labels and spatial or temporal boundaries. Source-level partitioning is used to reduce data leakage.

Experiments use an NVIDIA RTX 4090 workstation. Static detection uses 640×640 inputs; dynamic recognition uses 16 uniformly sampled frames; LoRA uses rank 8; and retrieval uses top-20 candidates followed by top-4 reranked passages. Static

TABLE I  
MAIN DATASET STATISTICS

| Dataset             | Samples | Categories | Test |
|---------------------|---------|------------|------|
| Static images       | 3,120   | 8          | 468  |
| Static instances    | 3,684   | 8          | —    |
| Dynamic clips       | 1,440   | 8          | 216  |
| Instructions        | 6,400   | —          | 640  |
| Knowledge documents | ~420    | —          | —    |
| Knowledge chunks    | ~3,200  | —          | —    |

recognition is evaluated using Precision, Recall, mAP@0.5, and mAP@0.5:0.95; dynamic recognition using Accuracy and Macro-F1; knowledge retrieval using Recall@K and MRR; and semantic interpretation using category accuracy, semantic similarity, evidence consistency, and hallucination rate. For open-ended semantic evaluation, cosine similarity between the reference and generated interpretations is computed using normalised sentence embeddings:

$$s(y, \hat{y}) = \frac{\mathbf{e}(y) \cdot \mathbf{e}(\hat{y})}{\|\mathbf{e}(y)\| \|\mathbf{e}(\hat{y})\|}. \quad (2)$$

A validation-tuned threshold  $\tau_s = 0.75$  determines semantic correctness. Responses below  $\tau_s$  are manually reviewed as hallucination candidates, and hallucination rate is computed as

$$HR = \frac{N_{\text{halluc}}}{N_{\text{total}}}, \quad (3)$$

where  $N_{\text{halluc}}$  denotes manually confirmed hallucinations.

## VII. RESULTS

### A. Static Cultural Symbol Detection

TABLE II  
STATIC CULTURAL SYMBOL DETECTION

| Model             | Prec.       | Recall      | mAP@.5      | mAP@.5:.95  |
|-------------------|-------------|-------------|-------------|-------------|
| YOLOv8s           | .868        | .842        | .887        | .632        |
| YOLOv10s          | .877        | .851        | .898        | .648        |
| YOLO11s           | .886        | .861        | .911        | .662        |
| DeepSymbol Static | <b>.905</b> | <b>.883</b> | <b>.934</b> | <b>.691</b> |

The DeepSymbol static branch reaches 0.934 mAP@0.5 and 0.691 mAP@0.5:0.95, improving over YOLO11s by 2.3 and 2.9 percentage points, respectively. The parameter count increases from approximately 9.4M to 13.8M, reflecting the additional feature capacity of the enhanced detector. Figure 2 visualises the baseline comparison,

All eight categories remain above 0.924 AP@0.5. Remaining errors concentrate among visually similar patterns such as cloud, lotus, peony, and meander/thunder.

### B. Dynamic Gesture Recognition

The dynamic branch reaches 0.889 Accuracy and 0.883 Macro-F1. Accuracy increases from 0.781 for the single-frame baseline to 0.878 with temporal Transformer modelling and then to 0.889 with the complete branch, indicating the contribution of

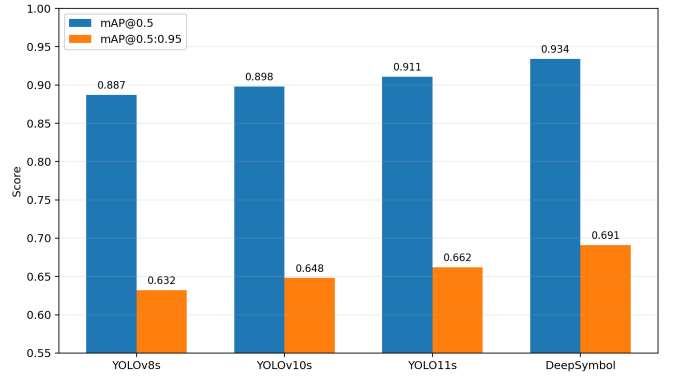


Fig. 2. Comparison of static cultural-symbol detection performance.

TABLE III  
DYNAMIC GESTURE RECOGNITION

| Model                | Accuracy    | Macro-F1    |
|----------------------|-------------|-------------|
| Single-frame         | .781        | .767        |
| CNN + LSTM           | .839        | .827        |
| CNN + GRU            | .848        | .835        |
| Temporal Transformer | .878        | .867        |
| DeepSymbol Dynamic   | <b>.889</b> | <b>.883</b> |

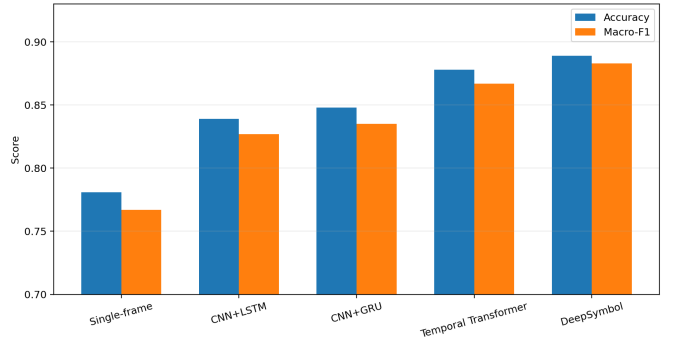


Fig. 3. Comparison of dynamic cultural-gesture recognition performance.

explicit temporal modelling. Figure 3 summarises the baseline comparison.

Representative qualitative outputs are shown in Fig. 4. The upper row contains three Double Happiness detection cases covering real-world scene complexity, multiple-instance detection, and intra-class variation. The lower rows contain representative temporal stages from G01 Fist-and-palm Salute, G02 Rock-on, and G03 Finger Heart. The original recognition overlays, class labels, confidence values, and frame indices are retained.

Remaining dynamic errors mainly occur between visually similar gesture pairs whose differences depend on movement direction and temporal progression.

### C. SwinV2 Ablation

To isolate the spatial-enhancement components, the YOLO11s configuration and training protocol are held constant

Static cultural-symbol detection



Dynamic gesture recognition

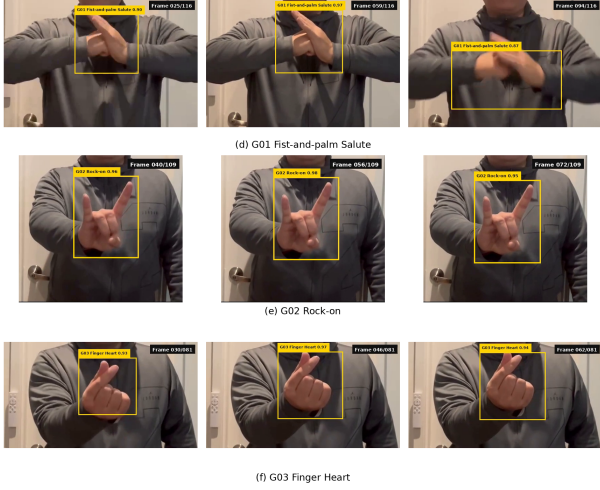


Fig. 4. Qualitative examples of static and dynamic cultural symbol recognition.

while window attention and shifted-window interaction are progressively introduced.

TABLE IV  
SWINV2 FEATURE-ENHANCEMENT ABLATION

| Configuration                | mAP@0.5 | mAP@0.5:0.95 | $\Delta$ mAP@0.5 |
|------------------------------|---------|--------------|------------------|
| YOLO11s baseline             | .911    | .662         | —                |
| + Window attention           | .919    | .670         | +0.008           |
| + Shifted-window interaction | .928    | .682         | +0.017           |
| Full SwinV2 enhancement      | .934    | .691         | +0.023           |

Window attention gives a 0.8-point gain, shifted-window interaction a cumulative 1.7-point gain, and the complete configuration a 2.3-point gain.

#### D. Temporal Sequence-Length Analysis

The temporal ablation shows that increasing the sequence length from 8 to 16 frames improves accuracy from 0.861 to 0.889, while 32 frames provides only a marginal further gain to 0.894.

#### E. Domain Adaptation

LoRA adaptation is evaluated against the base MLLM, few-shot prompting, and three LoRA ranks. The full results are shown in Table VI.

LoRA with  $r = 8$  increases category-answer accuracy from 0.684 to 0.816 and substantially improves format compliance and domain-appropriate expression. Increasing the rank to 16 provides only an additional 0.5 percentage points of category accuracy and a very small gain in the human rating. Considering

TABLE V  
TEMPORAL MODELLING AND SEQUENCE-LENGTH ABLATION

| Module               | Frames | Acc. | Macro-F1 | Cost        |
|----------------------|--------|------|----------|-------------|
| Mean pooling         | 16     | .816 | .802     | Low         |
| LSTM                 | 16     | .839 | .827     | Medium      |
| GRU                  | 16     | .848 | .835     | Medium      |
| Temporal Transformer | 8      | .861 | .844     | Medium      |
| Temporal Transformer | 16     | .889 | .883     | Medium-high |
| Temporal Transformer | 32     | .894 | .887     | High        |

TABLE VI  
LoRA DOMAIN-ADAPTATION RESULTS

| Method             | Acc. | Macro-F1 | Format | BERTScore | Prof. (1–5) |
|--------------------|------|----------|--------|-----------|-------------|
| Base MLLM          | .684 | .662     | .742   | .781      | 3.18        |
| Few-shot prompting | .721 | .704     | .834   | .803      | 3.46        |
| LoRA $r = 4$       | .783 | .768     | .922   | .836      | 3.88        |
| LoRA $r = 8$       | .816 | .804     | .961   | .857      | 4.16        |
| LoRA $r = 16$      | .821 | .807     | .963   | .860      | 4.18        |

trainable capacity, memory use, and small-sample overfitting risk,  $r = 8$  is retained. The result further indicates that LoRA principally adapts terminology, task formatting, and visual-language mapping rather than serving as a complete cultural knowledge repository.

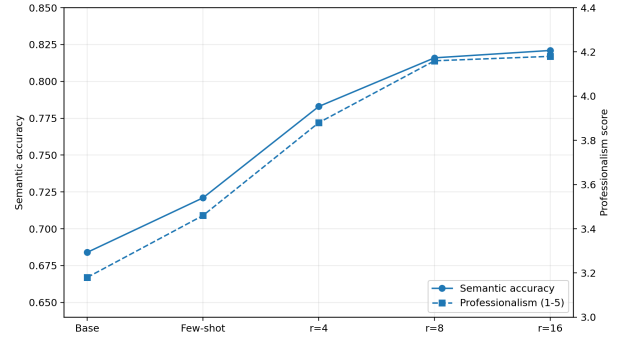


Fig. 5. Semantic adaptation performance under different LoRA configurations.

#### F. Knowledge Retrieval and Semantic Reasoning

Dense retrieval achieves Recall@20 of 0.913. Cross-encoder reranking improves MRR from 0.728 to 0.824 and Recall@5 from 0.788 to 0.831, mainly improving the rank of already retrieved evidence.

The unchanged Recall@20 confirms that reranking mainly improves evidence ordering rather than recall.

TABLE VII  
KNOWLEDGE-ENHANCED SEMANTIC REASONING

| Configuration     | Sem. Acc.   | Evid. Cons. | Prof. (1–5) | Halluc.     |
|-------------------|-------------|-------------|-------------|-------------|
| LoRA only         | .816        | .410        | 4.16        | 10.9%       |
| + Dense retrieval | .852        | .680        | 4.21        | 8.0%        |
| + Reranking       | .879        | .790        | 4.28        | 6.1%        |
| Full gated RAG    | <b>.893</b> | <b>.860</b> | <b>4.34</b> | <b>4.8%</b> |

The full RAG configuration improves semantic accuracy by 7.7 percentage points over LoRA-only adaptation. Evidence consistency increases from 0.410 to 0.860, while hallucination rate decreases from 10.9% to 4.8%. Figure 6 visualises this progressive improvement. These results demonstrate complementary roles for parameter adaptation and external evidence.

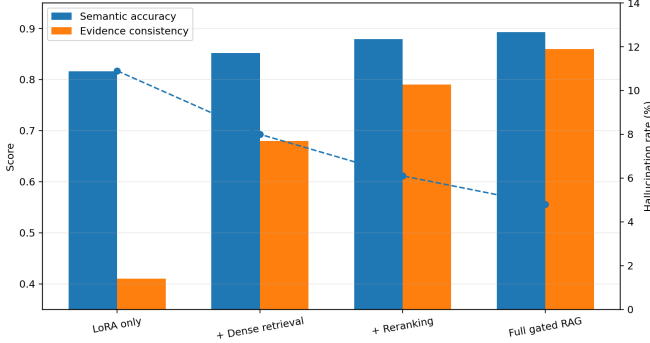


Fig. 6. RAG ablation results across three evaluation metrics.

### G. End-to-End Qualitative Analysis

Figure 7 shows the complete path from visual recognition to structured semantic interpretation and evidence grounding. The output separates directly observable visual information from broader cultural interpretation.

## VIII. DISCUSSION

Static and dynamic symbols benefit from different modelling strategies. Enhanced SwinV2 improves static mAP@0.5 by 2.3 points, while temporal modelling raises dynamic Accuracy from 0.781 to 0.878 before the complete branch reaches 0.889. Remaining errors mainly involve visually similar patterns or gestures.

For semantic interpretation, domain adaptation and evidence grounding address different sources of error. LoRA improves domain-specific task behaviour, while retrieval and evidence checking provide external support for cultural interpretations. The complete RAG configuration achieves 0.893 semantic accuracy and 0.860 evidence consistency, while reducing hallucination from 10.9% to 4.8%. This suggests that adaptation improves task-specific interpretation capability, whereas evidence grounding helps control unsupported responses.

The evaluation covers selected cultural symbols and gestures, so broader cross-source generalisation remains to be established. Semantic assessment also includes researcher-led evaluation, limiting its independence.

## IX. CONCLUSION

DeepSymbol integrates specialised spatial and temporal perception with task-specific semantic adaptation and evidence-aware reasoning for cultural symbol understanding. Across the evaluated tasks, the results indicate that these complementary components can support fine-grained recognition and more reliable semantic interpretation through domain adaptation and

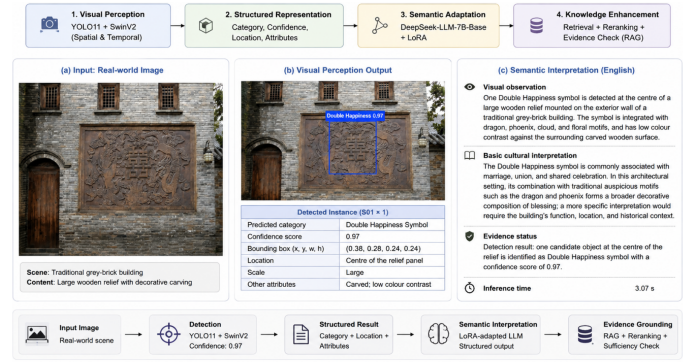


Fig. 7. End-to-end demonstration of DeepSymbol.

external evidence. Future work will extend the framework to broader cultural sources and more independent semantic evaluation.

## ACKNOWLEDGMENT

The authors used ChatGPT for language refinement and text condensation. All technical content and conclusions were developed and verified by the authors.

## REFERENCES

- [1] S. Nayak, K. Jain, R. Awal, S. Reddy, S. van Steenkiste, L. A. Hendricks, K. Stanczak, and A. Agrawal, "Benchmarking vision language models for cultural understanding," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 5769–5790.
- [2] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, S. Lin *et al.*, "Swin Transformer V2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 009–12 019.
- [3] K. Dong and W. Q. Yan, "Player performance analysis in table tennis through human action recognition," *Computers*, vol. 13, no. 12, p. 332, 2024.
- [4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 34 892–34 916.
- [5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.
- [6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [7] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5558–5570.
- [8] A. Alfarano, L. Venturoli, and D. N. Del Castillo, "VQArt-Bench: A semantically rich VQA benchmark for art and cultural heritage," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2025, pp. 396–406.
- [9] C. Miao, X. Wang, and W. Q. Yan, "Trademark detection and classification from digital images with self-attention mechanism," *Multimedia Tools and Applications*, vol. 85, no. 8, p. 633, 2026.
- [10] E. Liang, D. Wei, F. Li, H. Lv, and S. Li, "Object detection model of vehicle-road cooperative autonomous driving based on improved YOLO11 algorithm," *Scientific Reports*, vol. 15, p. 32348, 2025.
- [11] J. Qi, M. Nguyen, and W. Q. Yan, "CISO: Co-iteration semi-supervised learning for visual object detection," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 33 941–33 957, 2024.