From Separation to Entanglement: Dissecting Feed-Forward Network Functional Organization in Vision-Language Models for Ancient Chinese Documents

Ruilin Liu[1,2], Dongmei Zhu[1,2], Yuting Wei[3], Jin Song[4,5]*
[1]School of Information Management, Nanjing Agricultural University
[2]Research Center for Humanities and Social Computing, Nanjing Agricultural University
[3]School of Cyber Science and Engineering, University of International Relations
[4]China Mobile Jiutian Artificial Intelligence Technology (Beijing) Co., Ltd.
[5]School of Advanced Interdisciplinary Science, University of Chinese Academy of Sciences
liurl@stu.njau.edu.cn, songjinit@chinamobile.com

Abstract

While Vision-Language Models (VLMs) have demonstrated strong capabilities in general multimodal tasks, they face significant challenges in low-resource, cross-temporal settings such as ancient Chinese documents, where models must overcome complex visual noise and leverage deep linguistic priors. The functional organization of task-relevant components in existing VLMs remains understudied from a computational interpretability perspective. To address this, we introduce AncientVision-3T, an intervention-oriented diagnostic benchmark designed for mechanistic interpretability experiments on VLMs, organized across OCR, image classification, and image understanding tasks of increasing abstraction. Based on this benchmark, we propose the Ancient Chinese Target-Conditioned Gradient-based Neuron analysis framework (AC-TCGN), which identifies and validates task-relevant neurons for dissecting functional organization in general and domain-specific VLMs. Our experiments demonstrate that AC-TCGN effectively identifies neurons that are critical for model predictions. Moreover, our experimental results reveal a transition from separation to entanglement in the functional organization of decoder feed-forward networks: general VLMs show stronger task-wise separability, whereas the ancient Chinese domain-specific VLM shows greater cross-task neuron sharing and reduced task-wise separability. This suggests enhanced coupling between visual form and historical-semantic processing in ancient Chinese document image-text tasks.

1 Introduction

The rapid evolution of VLMs has revolutionized machine perception and understanding. Driven by their exceptional capabilities in multimodal tasks, there is growing interest in their potential application to low-resource and culturally specific domains (Dossou et al., 2025; Wei et al., 2024), particularly ancient Chinese documents. However, applying general visual models to these manuscripts, which embody millennia of historical culture, presents severe challenges. Rather than functioning as trivial multimodal composites, these documents constitute a dual chasm between visual perception and linguistic comprehension (Zhang et al., 2024b). The visual domain is complicated by vertical text flow and degradation noise caused by temporal erosion. Meanwhile, the linguistic interpretation demands deep priors to decode abstruse grammar and archaic glyphs. Consequently, the automated processing of ancient manuscripts extends beyond simple image recognition (Qiao et al., 2025), requiring the integration of visual feature extraction with higher-level semantic representations (Cao et al., 2025; Wang and Zhu, 2025).

Despite recent performance gains on ancient Chinese document image-text tasks, existing work remains largely outcome-oriented (Yang et al., 2024; Yu et al., 2025), with limited analysis of the internal processes underlying model predictions (Xu et al., 2025b; Wang et al., 2025b). It remains unclear how VLM parameters and neurons are functionally organized across tasks requiring different levels of visual perception and semantic reasoning, and whether neurons for visual recognition of ancient character forms and historical-semantic interpretation form separable or entangled patterns. This gap limits our understanding of domain adaptation in low-resource, cross-temporal settings and hinders the development of more interpretable and reliable multimodal systems.

To facilitate a structured analysis of decoder feed-forward network (FFN) neuron-level functional organization in VLMs on ancient Chinese

*Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 23176–23189
October 24-29, 2026 ©2026 Association for Computational Linguistics

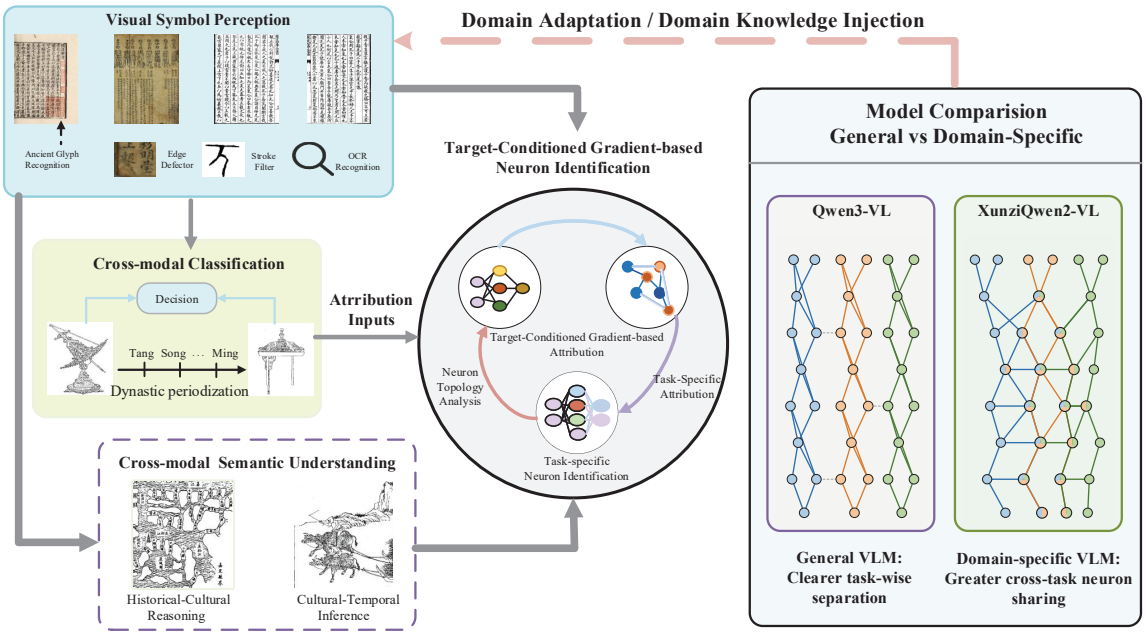


Figure 1: Task-wise separation and entanglement of decoder FFN neurons across general and ancient Chinese domain-specific VLMs.

document image–text tasks, we introduce a hierarchical diagnostic benchmark and interpretability framework, as illustrated in Figure 1. The benchmark organizes tasks into OCR, image classification, and image understanding, progressing from basic character recognition to visual categorization and ultimately to complex semantic reasoning. This design enables a systematic analysis of internal neuronal activation patterns across varying levels of semantic depth. Building upon this benchmark, we propose AC-TCGN, a target-conditioned neuron analysis framework designed for ancient Chinese image–text tasks, to conduct a comparative analysis between a general model and a domain-specific model. Focusing on the neuron-level analysis of FFN components in the Transformer decoder, we examine how domain specialization is associated with the distribution of task-relevant neurons and whether the domain-specific model exhibits increased FFN-level representational redundancy or reduced task separability compared with the general model.

In summary, the contributions of our work are as follows:

- We introduce AncientVision-3T, a hierarchical diagnostic benchmark for ancient Chinese document image-text tasks, designed to support controlled neuron-level interventions across OCR, image classification, and image understanding[1].
- We propose AC-TCGN, a target-conditioned neuron analysis framework that identifies and validates task-relevant FFN neurons to reveal the internal functional organization of VLMs for ancient Chinese documents.
- Through experiments on general VLMs and XunziQwen2-VL-7B, we find that the domain-specific model exhibits lower task separability and stronger cross-task interference at the decoder FFN neuron level in ancient Chinese document image-text tasks.

2 Related Work

Recent NLP research has increasingly focused on low-resource languages and cross-temporal corpora (Duan et al., 2024; Nacar et al., 2025; Yuan et al., 2025), which are fundamentally challenged by data scarcity, complex knowledge backgrounds (Do et al., 2025), and the difficulty of modeling historical semantics. However, benchmarks such as AncientDoc (Yu et al., 2025) and MCS-Bench (Liu et al., 2025) remain largely OCR-centric, despite their value for performance evaluation, and do not explicitly disentangle visual and semantic abstraction levels. This limits their support for fine-grained FFN neuron analysis and controlled interventions. AncientVision-3T addresses this gap with a compact, quality-controlled diagnostic benchmark featuring explicit task-level stratification. Similarly, on the methodological side,

[1]https://github.com/liu-ruilin/AncientVision-3T

methods have been explored, including contrastive distillation (Wei et al., 2025) and cross-modal fusion (Jiang et al., 2024). Although supervised fine-tuning and reinforcement learning have led to performance improvements (Li et al., 2025), most existing work remains predominantly performance-driven (Yao et al., 2025; Zhang et al., 2024a) and provides limited insight into internal model mechanisms.

Recent studies have increasingly focused on the internal mechanisms of VLMs. A corruption-based interpretability pipeline has been developed to analyze how visual information contributes to multimodal predictions (Golovanevsky et al., 2025), while distinct attention heads have been identified for visual text recognition (Baek et al., 2025). Modality-specific computational circuits involved in visual text recognition have also been investigated (Nikankin et al., 2026), alongside efforts to localize and edit multimodal neurons associated with visual and linguistic information (Pan et al., 2024). At the neuron level, prior work has characterized the functional roles of visual, textual, and multimodal neurons (Xu et al., 2025a) and examined how groups of neurons collectively contribute to multimodal concepts (Wang et al., 2025a). Despite this progress, the cross-task functional organization of task-relevant FFN neurons, particularly the extent to which these neurons are task-specific or shared in domain-specific VLMs, remains relatively underexplored.

3 AncientVision-3T

AncientVision-3T serves as a diagnostic benchmark targeted at mechanistic interpretability. Scaled specifically for fine-grained neuron-level analysis, it structures tasks across three hierarchical tiers of abstraction: OCR, image classification, and image understanding.

3.1 Data Resources

The images are primarily sourced from the Yinde Digital Humanities Platform, a digital archive integrating large-scale classical Chinese resources for historical and literary research. Built upon the China Biographical Database and supported by artificial intelligence technologies, the platform integrates large-scale classical Chinese resources for ancient literature research and historical data processing.

3.2 Data Construction

During the data collection process, we selected sources from the *Siku Quanshu* and various local gazetteers, spanning diverse historical periods including the Song, Yuan, Ming, and Qing dynasties, and encompassing both woodblock-printed editions and handwritten manuscripts.

We adopt a hybrid annotation pipeline in which Qwen-VL-Max generates preliminary annotations and five experts in classical Chinese philology and literature determine the final labels. Each instance is verified by one expert and reviewed by a second, with disagreements adjudicated by a third. For OCR, annotators preserve reading order, rare and variant characters, and layout structure; classification labels must follow predefined class definitions and visible evidence; image-understanding annotations are checked for visual grounding, historical-semantic correctness, completeness, and unsupported content. Duplicate, illegible, severely degraded, or image–text-mismatched instances are excluded.

For inter-annotator reliability, two experts independently annotated a blinded, stratified subset of 300 instances (100 per task), covering relevant dynasties, media types, classes, and semantic categories. Before adjudication, inter-annotator agreement was 0.984 as measured by character-level $1 - \mathrm{NED}$ for OCR, Cohen's $\kappa = 0.972$ for classification, and Krippendorff's $\alpha = 0.905$ for image understanding. These scores indicate high agreement across all three tasks. Remaining disagreements were subsequently resolved by a third expert.

3.3 Data Statistics

The dataset comprises two distinct parts: textual images and illustrative images. The textual subset consists of 500 images dedicated to OCR tasks, featuring Traditional Chinese text primarily arranged in a vertical text flow. The illustration subset consists of 1,000 images. The classification task focuses on objects with explicit cultural semantics, such as ritual vessels, musical instruments, official attire, and military equipment, while the image understanding task spans a total of 20 categories, including architecture, hydraulic engineering, astronomy and calendrical systems, and cartography.

4 Methodology

Following prior neuron-level interpretability studies (Dai et al., 2022), we focus on FFN neurons

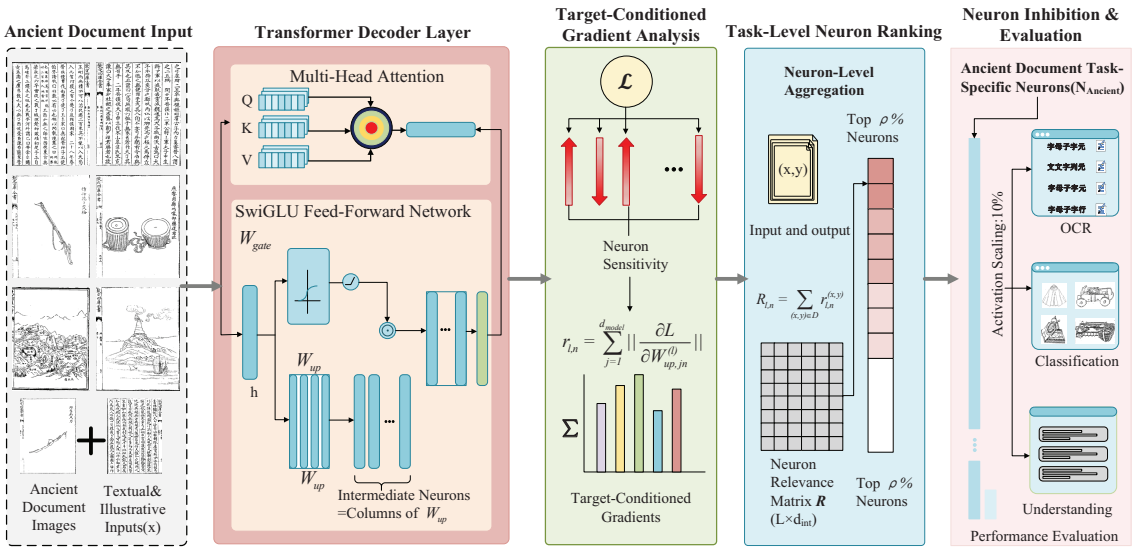


Figure 2: Overview of the AC-TCGN framework for identifying and validating task-relevant FFN neurons in VLMs.

in the Transformer decoder. FFNs have been associated with semantic information processing in Transformer models (Geva et al., 2021), while the decoder provides the stage at which visual representations interact with linguistic context. As reported in Section C, our centered kernel alignment (CKA) analysis of the visual encoder further shows high cross-task representational similarity across OCR, image classification (IC), and image understanding (IU), providing additional motivation for analyzing task-level functional organization in decoder FFNs.

4.1 Background

Modern LLMs and VLMs are primarily built upon the Transformer decoder architecture (Vaswani et al., 2017), where two components play a central role: multi-head self-attention and the feed-forward network.

Given the hidden state $\boldsymbol{h}^{i-1} \in \mathbb{R}^d$ from the $(i-1)$-th layer, the multi-head attention (MHA) module updates the representation as:

$$\tilde{\boldsymbol{h}} = \mathrm{Attn}(\boldsymbol{h}^{i-1}\boldsymbol{W}_q^i, \boldsymbol{h}^{i-1}\boldsymbol{W}_k^i, \boldsymbol{h}^{i-1}\boldsymbol{W}_v^i)\boldsymbol{W}_o^i, \tag{1}$$

where $\boldsymbol{W}_q^i$, $\boldsymbol{W}_k^i$, and $\boldsymbol{W}_v^i$ denote the weight matrices for queries, keys, and values, respectively, and $\boldsymbol{W}_o^i$ is the output weight matrix. $\boldsymbol{h}^{i-1}$ represents the output from the previous layer.

Following the attention update, each Transformer layer applies a position-wise FFN to enrich non-linear transformations.

For models using the standard two-layer FFN, the computation can be written as:

$$\boldsymbol{h}^i = \mathrm{act\_fn}(\tilde{\boldsymbol{h}}^i\boldsymbol{W}_1^i)\boldsymbol{W}_2^i, \tag{2}$$

where $\boldsymbol{W}_1^i \in \mathbb{R}^{d\times 4d}$ and $\boldsymbol{W}_2^i \in \mathbb{R}^{4d\times d}$, and $\mathrm{act\_fn}(\cdot)$ is an activation function such as GELU (Hendrycks, 2016).

4.2 Neuron Identification for Ancient Chinese Document Image-Text Tasks

Ancient Chinese document image-text tasks require fine-grained cross-modal reasoning, where predictions depend on subtle visual evidence aligned with historically grounded semantics. Unlike modern multimodal tasks, understanding in this domain is less driven by prompt templates or special tokens (Fang et al., 2024; Pan et al., 2024), but by strong visual-target dependencies. We therefore identify task-relevant neurons from a target-conditioned gradient perspective, scoring each decoder FFN up-projection neuron by its sensitivity to task-specific supervision. Figure 2 illustrates AC-TCGN as a task-level framework that integrates neuron scoring, global ranking, and causal inhibition validation.

4.2.1 Neuron Definition in SwiGLU Feed-Forward Networks

We focus on FFNs in Transformer decoders, which have been shown to facilitate high-level and task-related semantic transformations. In Qwen-based VLMs, FFNs adopt the SwiGLU architecture, defined at the l-th layer as:

$$\mathrm{FFN}^{(l)}(\tilde{\boldsymbol{h}}) = (\mathrm{SiLU}(\boldsymbol{h}\mathbf{W}_{gate}^{(l)}) \odot (\boldsymbol{h}\mathbf{W}_{up}^{(l)}))\mathbf{W}_{down}^{(l)} \tag{3}$$

where $\boldsymbol{W}_{gate}^{(l)}$, $\boldsymbol{W}_{up}^{(l)}$, and $\boldsymbol{W}_{down}^{(l)}$ denote the gating, up-projection, and down-projection matrices, respectively.

Following prior neuron-level interpretability studies, we define each column of the up-projection matrix $\boldsymbol{W}_{up}^{(l)}$ as an intermediate neuron. These neurons constitute the fundamental functional units through which visual and semantic information is transformed within the model.

4.2.2 Target-Conditioned Gradient-Based Neuron Identification

Given a dataset for ancient Chinese document image-text tasks $\mathcal{D} = (x, y)$, where x consists of ancient images and associated textual context and y denotes the target outputs, we compute gradients with respect to the up-projection parameters for each sample.

Specifically, for the l-th layer, we compute:

$$\boldsymbol{G}_{up}^{(l)} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{up}^{(l)}}, \tag{4}$$

where $\mathcal{L}$ denotes the task loss computed on ground-truth target outputs and serves as the supervision signal for ancient Chinese document image-text tasks such as OCR, classification, and semantic interpretation.

We then define the identification score of the n-th intermediate neuron in layer l as the l_1-norm of the gradient of the n-th column of the up-projection matrix:

$$r_{l,n} = \| \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{up}^{(l)}[:, n]} \|_1 = \sum_{j=1}^{d_{\text{model}}} \| \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{up,jn}^{(l)}} \|, \tag{5}$$

where n indexes intermediate neurons in the up-projection matrix and d_{model} denotes the dimensionality of the model's hidden states.

This score measures the sensitivity of the task loss to perturbations of the parameters associated with a given neuron and serves as a criterion to identify neurons that are locally influential to the target task. Calculating gradients for $\boldsymbol{W}_{up}$ remains consistent with the SwiGLU formulation, as $\boldsymbol{W}_{up}$ linearly scales the content term prior to gating. Under a first-order approximation, these perturbations propagate approximately linearly, establishing a principled link between gradient-based identification and activation-level intervention.

To stably identify ancient Chinese document image-text task neurons, we aggregate the identification scores across the dataset:

$$R_{l,n} = \sum_{(x,y)\in\mathcal{D}} r_{l,n}^{(x,y)}, \tag{6}$$

where $r_{l,n}^{(x,y)}$ denotes the neuron identification score computed for sample (x, y). This aggregation yields a task-level neuron relevance matrix:

$$R \in \mathbb{R}^{L \times d_{\text{int}}}, \tag{7}$$

where L denotes the number of Transformer layers and d_{int} denotes the intermediate FFN dimension. Finally, we rank all neurons globally and select the top $\rho\%$ neurons with the highest relevance scores as the ancient Chinese document task neuron set:

$$\mathcal{N}_{\text{Ancient}} = \text{Top}_{\rho}\left(\{R_{l,n} \mid \forall l, n\}\right). \tag{8}$$

5 Experiments

To investigate FFN neuron-level functional organization in VLMs for ancient Chinese document image–text tasks, we conduct controlled intervention experiments on the AncientVision-3T diagnostic benchmark. We analyze two models: Qwen3-VL-8B, a general VLM, and XunziQwen2-VL-7B[2], a domain-specific model specialized for ancient Chinese corpora, across OCR, IC, and IU.

5.1 Dataset

For interpretability experiments, we split AncientVision-3T into training and validation sets with a 1:1 ratio. The training split serves only as an identification split for collecting neuron statistics and constructing intervention masks; no model parameters are updated. Inhibition effects are evaluated independently on the held-out validation split, following the controlled-data paradigm commonly used in neuron-level interpretability studies (Dai et al., 2022; Tang et al., 2024).

5.2 Experimental Setup

Decoding settings. We set the maximum generation length to 512 for OCR and IU and 32 for IC. All models are required to generate full textual outputs. Decoding is performed with temperature 0 and no sampling for deterministic generation. For neuron analysis, we select the top 5% of neurons as ancient Chinese document image-text task neurons and apply strong inhibition by scaling their activations to 10% of the original values, with the choice of these hyperparameters justified through sensitivity analysis in Section A.

[2]https://modelscope.cn/models/XunziALLM/Xunzi_VL

Method	Qwen3-VL-8B			XunziQwen2-VL-7B		
	OCR Char. Acc.(%)	IC Acc.(%)	IU Rouge-L(%)	OCR Char. Acc.(%)	IC Acc.(%)	IU Rouge-L(%)
Baseline	87.94	78.00	8.06	97.57	47.20	23.24
RS	$87.30_{(-0.64)}$	$77.20_{(-0.80)}$	$9.39_{(+1.33)}$	$95.57_{(-2.00)}$	$46.00_{(-1.20)}$	$22.93_{(-0.31)}$
ActGrad	$22.58_{(-65.36)}$	$50.80_{(-27.20)}$	$5.52_{(-3.54)}$	$54.03_{(-43.54)}$	$30.40_{(-16.80)}$	$5.95_{(-17.29)}$
LAP	$10.98_{(-76.96)}$	$61.60_{(-16.40)}$	$5.78_{(-2.28)}$	$65.43_{(-32.14)}$	$32.80_{(-14.40)}$	$21.69_{(-1.55)}$
LAPE	$13.80_{(-74.14)}$	$58.80_{(-19.20)}$	$4.73_{(-3.33)}$	$53.74_{(-43.83)}$	$44.40_{(-2.80)}$	$13.34_{(-9.90)}$
AC-TCGN	$\mathbf{2.98}_{(-84.96)}$	$\mathbf{29.60}_{(-48.40)}$	$\mathbf{3.26}_{(-4.80)}$	$\mathbf{19.83}_{(-77.74)}$	$\mathbf{42.40}_{(-4.80)}$	$\mathbf{5.62}_{(-17.62)}$

Table 1: Baseline performance and performance changes after inhibiting target-task neurons identified by different methods. The baseline row reports the absolute performance of each VLM, while the remaining rows report performance changes relative to the baseline after inhibiting target-task neurons identified by different methods. Char. Acc. denotes character-level accuracy, computed as 1-NED.

Identification methods. We compare our method with several neuron-scoring baselines under the same task-level aggregation and global top-$\rho\%$ selection pipeline, so that observed differences isolate the contribution of the scoring function itself. Random Selection (RS) samples the same number of intermediate neurons as a control. ActGrad (Leng and Xiong, 2025) scores neurons via a first-order Taylor approximation using activation–gradient products. LAP (Dai et al., 2022; Gurnee et al., 2024; Voita et al., 2024) selects neurons based on their activation probability on target-task samples, while LAPE (Tang et al., 2024) ranks neurons by the entropy of activation probabilities across tasks.

Metrics. For the three tasks considered in this study, we adopt different evaluation metrics. Since OCR on ancient Chinese document images typically emphasizes character-level correctness, we use Normalized Edit Distance (NED) as the evaluation metric (Levchenko, 2025), where 1-NED corresponds to the character-level accuracy. IC tasks are evaluated using accuracy. For IU, we adopt Rouge-L as the evaluation metric.

6 Results

6.1 Effectiveness of Neuron Identification

We first evaluate the baseline performance of Qwen3-VL-8B and XunziQwen2-VL-7B on AncientVision-3T. As shown in Table 1, XunziQwen2-VL-7B outperforms Qwen3-VL-8B on OCR and IU, indicating stronger capability in ancient Chinese document processing. We then evaluate the inhibition effects of neurons identified by AC-TCGN and the baseline methods. Overall, AC-TCGN induces the most pronounced performance degradation in most settings, suggesting that its identified neuron sets have stronger functional relevance to task outputs. The inhibition patterns further reveal differences in FFN-level functional organization across models and tasks. In particular, the domain-specific model exhibits greater sensitivity in image understanding, suggesting a closer association between visual evidence and historical-semantic processing in ancient Chinese document tasks.

6.2 Cross-Task Performance and Analysis

If task-relevant neurons are shared across tasks, inhibiting neurons for a target task t may also affect other tasks $s \neq t$, making cross-task performance degradation an empirical indicator of FFN neuron-level entanglement. Figure 3 illustrates the cross-task performance degradation induced by different neuron identification methods. On Qwen3-VL-8B, AC-TCGN demonstrates superior task selectivity and stability. In contrast, ActGrad exhibits moderate capability: while inhibiting OCR neurons causes less interference than LAP and LAPE, targeting IC and IU neurons triggers substantial cross-task degradation. Furthermore, LAP and LAPE induce synchronous performance fluctuations across tasks, reflecting unstable identification and limited task isolation. On XunziQwen2-VL-7B, AC-TCGN continues to substantially degrade target performance while causing noticeable cross-task interference, whereas other methods yield limited impact. These results suggest that XunziQwen2-VL-7B relies on more distributed and shared FFN neuron-level representations, as evidenced by the stronger cross-task interference observed under neuron inhibition.

6.3 Neuron Visualization and Analysis

We visualize the layer-wise activation patterns of neurons associated with OCR, IC, and IU. As shown in Figure 4, OCR-related neurons are con-

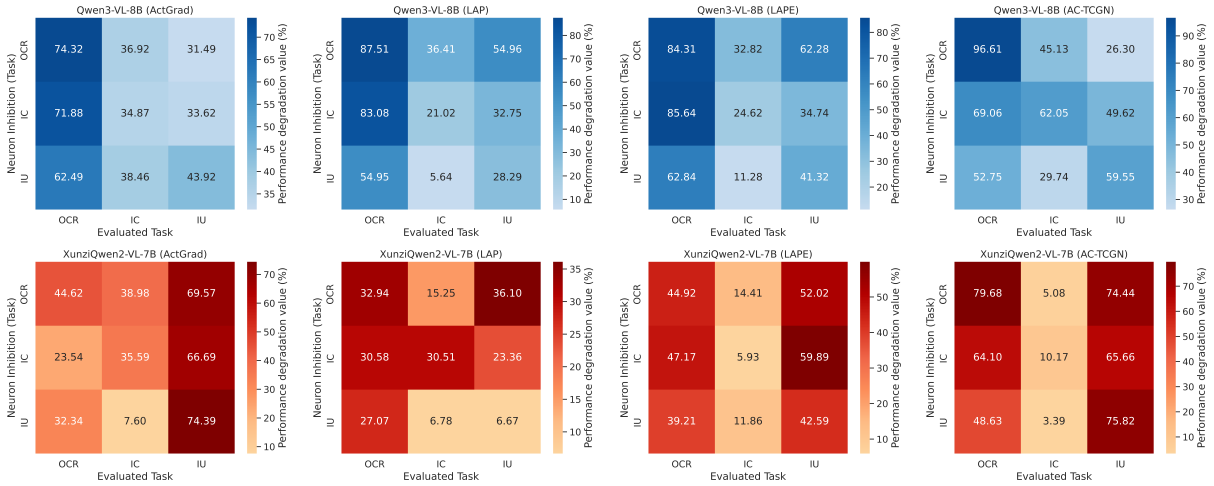


Figure 3: Heatmaps of cross-task effects induced by different neuron identification methods. On the y-axis, we show the target task whose neurons are identified and inhibited, and on the x-axis the evaluated task. Each cell reports the degree of performance degradation on the corresponding task after neuron inhibition, with diagonal cells indicating self-task inhibition. Results for Qwen3-VL-8B and XunziQwen2-VL-7B are shown in blue and red, respectively. The degree of performance degradation is computed as $(\text{Performance}_{\text{baseline}} - \text{Performance}_{\text{inhibited}})/\text{Performance}_{\text{baseline}}$.

centrated primarily in layers 4–7. XunziQwen2-VL-7B exhibits denser activation patterns in these layers than Qwen3-VL-8B, suggesting greater involvement of intermediate layers in text recognition. Layers 4–7 contain 33.1% and 36.5% of the OCR neurons identified in Qwen3-VL-8B and XunziQwen2-VL-7B, respectively, corresponding to approximately 3.0× and 2.6× the proportions expected under a uniform layer distribution. For IC, Qwen3-VL-8B exhibits a stronger concentration of task-related neurons in the upper layers: 63.7% of the identified IC neurons lie within the upper third of the model depth, compared with 30.4% in XunziQwen2-VL-7B. For IU, XunziQwen2-VL-7B shows a broader distribution across layers, whereas Qwen3-VL-8B exhibits a stronger concentration in the lower layers; 71.6% of the identified IU neurons lie within the lower third of Qwen3-VL-8B, compared with 48.1% in XunziQwen2-VL-7B. Beyond these layer-wise distributions, we quantify cross-task neuron sharing through pairwise Jaccard overlap among the AC-TCGN top-ρ neuron sets. XunziQwen2-VL-7B exhibits higher overlap than Qwen3-VL-8B, with OCR–IC, OCR–IU, and IC–IU overlaps of 10.3%, 20.6%, and 13.8%, respectively, compared with 6.4%, 15.7%, and 6.1% for Qwen3-VL-8B. In general, these overlap patterns provide complementary evidence of greater cross-task neuron sharing in XunziQwen2-VL-7B, consistent with stronger cross-task interference observed under neuron inhibition.

6.4 Cross-Task Selectivity Across Model Scales and Architectures

To quantify task separability and cross-task dynamics, we use Self-Task Drop, Cross-Task Drop, and the Separation Index (SI). As shown in Table 2, SI increases monotonically with model size within the Qwen3-VL family, suggesting clearer task-wise separability in larger architectures. Across architectures, all general-purpose models achieve higher SI values than XunziQwen2-VL-7B, which records the lowest SI of 1.27. This low SI is not due to an unusually large Cross-Task Drop, since its Cross-Task Drop of 43.55 falls within the range of general models, from 41.55 to 46.79. Instead, it mainly results from the lowest Self-Task Drop of 55.22. Thus, inhibiting task-specific neurons in XunziQwen2-VL-7B causes less additional target-task degradation while producing comparable collateral effects on other tasks, leading to the smallest margin between self-task and cross-task effects. This suggests that task-relevant computation in XunziQwen2-VL-7B is less concentrated within a compact, task-exclusive set of FFN neurons, consistent with the reduced separability observed in Figure 3.

6.5 Controlling for Model Generation Effects

As shown in Table 3, Qwen2-VL-7B exhibits a higher SI value than XunziQwen2-VL-7B, together with notably weaker cross-task interference. To further validate this observation, we conduct an

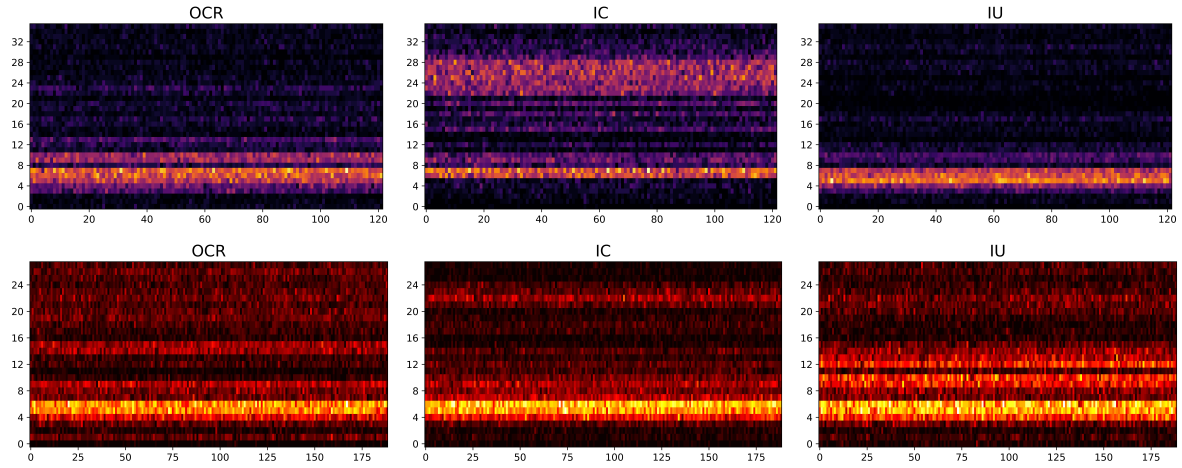


Figure 4: Visualization of neurons across three tasks: OCR, Image Classification (IC), and Image Understanding (IU). The y-axis denotes the task associated with the identified neurons whose activations are visualized, while the x-axis represents the model layer index. The top row corresponds to Qwen3-VL-8B, and the bottom row corresponds to XunziQwen2-VL-7B.

Model	OCR Acc. (%)	IC Acc. (%)	IU (%)	Self-Drop (%)	Cross-Drop (%)	SI (Self/Cross)
Qwen3-VL-4B	$2.37_{(-74.24)}$	$27.20_{(-49.20)}$	$4.48_{(-8.37)}$	74.50	46.79	1.59
Qwen3-VL-8B	$2.98_{(-84.96)}$	$29.60_{(-48.40)}$	$3.26_{(-4.80)}$	72.74	45.43	1.60
Qwen3-VL-32B	$0.16_{(-80.39)}$	$28.80_{(-54.40)}$	$4.54_{(-2.63)}$	76.16	45.61	1.67
MiniCPM-V-4.5	$5.72_{(-61.06)}$	$30.80_{(-47.60)}$	$6.06_{(-9.11)}$	69.39	41.55	1.67
LLaVA-OneVision-1.5-8B	$0.15_{(-6.44)}$	$8.00_{(-17.60)}$	$1.39_{(-3.42)}$	79.18	43.25	1.83
InternVL3.5-8B	$10.78_{(-79.48)}$	$25.20_{(-32.00)}$	$7.74_{(-9.10)}$	65.34	42.43	1.54
XunziQwen2-VL-7B	$19.83_{(-77.74)}$	$42.40_{(-4.80)}$	$5.62_{(-17.62)}$	55.22	43.55	1.27

Table 2: Cross-task inhibition and selectivity analysis across model scales and families. We report three metrics to quantify task selectivity. Self-Task Drop is the average relative performance degradation on the target task after inhibiting its task-specific neurons: $\frac{1}{3}\sum_t(\text{Perf}^{(t,t)}_{\text{baseline}} - \text{Perf}^{(t)}_{\text{inhibited}})/\text{Perf}^{(t)}_{\text{baseline}}$. Cross-Task Drop measures degradation on other tasks when inhibiting neurons of task t: $\frac{1}{6}\sum_{t\neq s}\text{degradation}(t \to s)$. Separation Index (SI) $= \frac{\text{Self-Task Drop}}{\text{Cross-Task Drop}}$, where larger values indicate clearer functional separation between tasks.

Model	OCR (%)	IC (%)	IU (%)	SI
Qwen2-VL-7B	$2.75_{(-47.29)}$	$25.20_{(-51.60)}$	$1.44_{(-3.37)}$	1.76
XunziQwen2-VL-7B	$19.83_{(-77.74)}$	$42.40_{(-4.80)}$	$5.62_{(-17.62)}$	1.27

Table 3: Performance comparison on OCR, IC, and IU tasks. The subscript indicates the relative performance drop compared with the baseline. SI denotes the separation index between self-task and cross-task interference.

additional analysis on Qwen2-VL-7B, a general-purpose VLM from the same model generation as XunziQwen2-VL-7B, following the same experimental protocol as in the main experiments. The neuron identification and cross-task interference results under AC-TCGN inhibition are presented in Section B. Despite belonging to the same model generation, the two models exhibit substantially different levels of cross-task interference. These results suggest that the stronger FFN neuron-level cross-task interference observed in XunziQwen2-VL-7B is unlikely to be explained solely by model generation differences and may instead be associated with its specialization for the ancient Chinese domain.

6.6 Generalizability Across Domain-Specific Models and Corpora

We examine the generalizability of the observed reduction in FFN-level task separability across domain-specific models and corpora. For model-level validation, we apply the same AC-TCGN protocol to TongGu-VL-2B-Instruct [3], another domain-specific VLM for classical Chinese, and compare it with Qwen2-VL-2B-Instruct. As shown in Table 5, TongGu-VL-2B-Instruct exhibits a lower SI than Qwen2-VL-2B-Instruct, with values of 1.43 and 1.55, respectively, consistent with the pattern observed for XunziQwen2-VL-7B. For corpus-level validation, we further evaluate 150 instances from

[3] https://github.com/SCUT-DLVCLab/TongGu-VL

Model	Inhibited	OCR Acc. (%)	Trans. ROUGE-L (%)	QA ROUGE-L (%)	SI
Qwen3-VL-8B	Baseline	67.55	16.19	9.03	1.82
	OCR	5.83	5.68	6.78	
	Trans	24.86	2.36	6.26	
	QA	37.06	9.65	2.93	
XunziQwen2-VL-7B	Baseline	92.36	24.31	20.95	1.39
	OCR	15.48	7.67	7.56	
	Trans	34.94	4.78	9.41	
	QA	49.48	12.33	4.77	

Table 4: Performance changes under task-specific neuron inhibition on the AncientDoc benchmark. Each row reports model performance when neurons associated with a particular task (OCR, Translation, or QA) are inhibited. SI denotes the separation index between self-task and cross-task interference.

Model	OCR (%)	IC (%)	IU (%)	SI
Qwen2-VL-2B-Instruct	$0.97_{(-37.84)}$	$19.60_{(-31.20)}$	$1.79_{(-6.17)}$	1.55
TongGu-VL-2B-Instruct	$14.74_{(-63.88)}$	$30.40_{(-6.80)}$	$3.28_{(-2.34)}$	1.43

Table 5: Matched comparison between Qwen2-VL-2B-Instruct and TongGu-VL-2B-Instruct under AC-TCGN inhibition.

AncientDoc[4], covering page-level OCR, vernacular translation, and knowledge-based QA. As shown in Table 4, XunziQwen2-VL-7B again exhibits lower task separability than Qwen3-VL-8B, with SI values of 1.39 and 1.82, respectively, while also showing stronger cross-task interference. Taken together, these results indicate that the observed reduction in FFN-level task separability is not specific to XunziQwen2-VL-7B and is not limited to AncientVision-3T.

7 Conclusion

This study introduces AncientVision-3T, a hierarchical diagnostic benchmark for ancient Chinese document image-text tasks, and a target-conditioned neuron analysis framework that combines task-level neuron identification with intervention-based validation. Through analyses of general VLMs and XunziQwen2-VL-7B, we examine differences in the functional organization of decoder FFN neurons. General VLMs exhibit clearer task-wise separability, whereas the domain-specific model shows greater cross-task neuron sharing and reduced task-wise separability across OCR, image classification, and image understanding. These findings may inform model development by revealing capability-specific neuron distributions and intervention effects, thereby guiding targeted adjustments to training data mixtures or auxiliary objectives. They may further support the assessment of unintended effects introduced by neuron-level editing, the design of domain adaptation strategies that preserve functional specialization, and the evaluation of whether different capabilities depend on shared or distinct computational pathways.

Limitations

Our analysis focuses primarily on FFN neurons in the Transformer decoder. These neurons provide an interpretable and intervention-friendly lens for examining task-relevant functional organization and perceptual-semantic associations, but they do not capture all internal components of VLMs. Our conclusions are also derived from image-text processing tasks on ancient Chinese documents, a domain characterized by low-resource supervision, cross-temporal linguistic variation, degraded visual inputs, and tight coupling between visual form and historical semantics. The observed separation-to-entanglement pattern should therefore be interpreted within this domain context. Further validation is needed to determine whether these conclusions generalize to other multimodal domains. Future work will extend the analysis to broader application scenarios and additional model components to provide a more comprehensive understanding of functional organization in VLMs.

Ethical Statement

The dataset used in this study is derived from publicly available digital archives of ancient Chinese documents, and all data usage complies with relevant copyright laws and academic licenses. We respect the cultural and historical value of ancient documents and ensure that our research methods and analyses follow established academic and ethical standards. In constructing and using

[4] https://github.com/bytedance/AncientDoc

AncientVision-3T, we pay particular attention to potential historical biases and cultural misunderstandings, so as to avoid reinforcing incorrect or outdated societal perspectives in content related to ancient Chinese documents. All models employed are open-source and used in accordance with their respective licenses, and this work aims to support the responsible application of AI technologies in low-resource language processing and understanding.

Acknowledgments

This work was supported by the National Social Science Foundation of China (Grants Nos. 21 and ZD331). We also acknowledge the support of the High-Performance Computing Public Platform of Nanjing Agricultural University. We thank Professor Dongbo Wang for his support, and Dr. Pengcheng Huang and Dr. Shiqi Chen for their valuable guidance.

References

Ingeol Baek, Hwan Chang, Sunghyun Ryu, and Hwanhee Lee. 2025. How do large vision-language models see text in image? unveiling the distinctive role of OCR heads. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20452–20464, Suzhou, China. Association for Computational Linguistics.

Jiahuan Cao, Yang Liu, Peirong Zhang, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. Tonggu-vl: Advancing visual-language understanding in chinese classical studies through parameter sensitivity-guided instruction tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11111–11120.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Thao Do, Dinh Phu Tran, An Vo, and Daeyoung Kim. 2025. Reference-based post-ocr processing with llm for precise diacritic text in historical document recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 27951–27959.

Bonaventure FP Dossou, Inès Arous, and Jackie CK Cheung. 2025. Rethinking full finetuning from pretraining checkpoints in active learning for african languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 64–78.

Siyu Duan, Jun Wang, and Qi Su. 2024. Restoring ancient ideograph: a multimodal multitask neural network approach. *arXiv preprint arXiv:2403.06682*.

Junfeng Fang, Zac Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. 2024. Towards neuron attributions in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:122867–122890.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. 2025. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11462–11482, Albuquerque, New Mexico. Association for Computational Linguistics.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*.

D Hendrycks. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hanqi Jiang, Yi Pan, Junhao Chen, Zhengliang Liu, Yifan Zhou, Peng Shu, Yiwei Li, Huaqin Zhao, Stephen Mihm, Lewis C Howe, et al. 2024. Oraclesage: Towards unified visual-linguistic understanding of oracle bone scripts through cross-modal knowledge fusion. *arXiv preprint arXiv:2411.17837*.

Yongqi Leng and Deyi Xiong. 2025. Towards understanding multi-task learning (generalization) of LLMs via detecting and exploring task-specific neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2969–2987, Abu Dhabi, UAE. Association for Computational Linguistics.

Maria Levchenko. 2025. Evaluating llms for historical document ocr: A methodological framework for digital humanities. *arXiv preprint arXiv:2510.06743*.

Shen Li, Renfen Hu, and Lijun Wang. 2025. Efficiently building a domain-specific large language model from scratch: A case study of a classical chinese large language model. *arXiv preprint arXiv:2505.11810*.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. MCS-bench: A comprehensive benchmark for evaluating multimodal large language models in Chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10435–10492, Vienna, Austria. Association for Computational Linguistics.

Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, et al. 2025. Towards inclusive arabic llms: A culturally aligned benchmark in arabic large language model evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. 2026. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *Advances in Neural Information Processing Systems*, 38:66352–66385.

Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. 2024. Finding and editing multi-modal neurons in pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1012–1037, Bangkok, Thailand. Association for Computational Linguistics.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Jiapeng Wang, YiFan Zhang, Zhuoma GongQue, Chong Sun, Yida Xu, Yadong Xue, Ye Tian, Zhimin Bao, Lan Yang, Chen Li, and Honggang Zhang. 2025. V-oracle: Making progressive reasoning in deciphering oracle bones for you and me. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20124–20150, Vienna, Austria. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301.

Dongbo Wang and Dongmei Zhu. 2025. Benchmarking the ancient books capability of multimodal large language models. *npj Heritage Science*, 13(1):339.

Feiyu Wang, Ziran Zhao, Dong Yu, and Pengyuan Liu. 2025a. Attribution and application of multiple neurons in multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11649–11662, Suzhou, China. Association for Computational Linguistics.

Haofeng Wang, Yilin Guo, Tiange Zhang, Zehao Li, Tong Yue, Yizong Wang, Rongqun Lin, Feng Gao, Shiqi Wang, and Siwei Ma. 2025b. Riverecho-2.0: A real-time interactive system for yellow river culture via enhanced multimodal document rag. *Transactions on Artificial Intelligence*, 1(1):212–226.

Yuting Wei, Qi Meng, Yuanxing Xu, and Bin Wu. 2025. TEACH: A contrastive knowledge adaptive distillation framework for classical Chinese understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3537–3550, Vienna, Austria. Association for Computational Linguistics.

Yuting Wei, Yuanxing Xu, Xinru Wei, Yangsimin Yangsimin, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. Ac-eval: Evaluating ancient chinese language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1600–1617.

Jiaqi Xu, Cuiling Lan, and Yan Lu. 2025a. Deciphering functions of neurons in vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3173–3181.

Zishan Xu, Xiaofeng Zhang, Yuqing Yang, Wei Chen, Jueting Liu, Tingting Xu, Zehua Wang, and Abdulmotaleb El Saddik. 2025b. Muralagent: Enhancing ancient mural outpainting with rag-based texts and multimodal integration. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(9):1–17.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. 2024. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*.

Xinyu Yao, Mengdi Wang, Bo Chen, and Xiaobing Zhao. 2025. Wenyangpt: A large language model for classical chinese tasks. *arXiv preprint arXiv:2504.20609*.

Haiyang Yu, Yuchuan Wu, Fan Shi, Lei Liao, Jinghui Lu, Xiaodong Ge, Han Wang, Minghan Zhuo, Xuecheng Wu, Xiang Fei, et al. 2025. Benchmarking vision-language models on chinese ancient documents: From ocr to knowledge reasoning. *arXiv preprint arXiv:2509.09731*.

Junyi Yuan, Jian Zhang, Fangyu Wu, Huanda Lu, Dongming Lu, and Qiufeng Wang. 2025. Towards cross-modal retrieval in chinese cultural heritage documents: Dataset and solution. In *International Conference on Document Analysis and Recognition*, pages 570–586. Springer.

Yuqing Zhang, Baoyi He, Yihan Chen, Hangqi Li, Han Yue, Shengyu Zhang, Huaiyong Dou, Junchi Yan, Zemin Liu, Yongquan Zhang, et al. 2024a. Philogpt: A philology-oriented large language model for ancient chinese manuscripts with dunhuang as case study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2784–2801.

Yuqing Zhang, Hangqi Li, Shengyu Zhang, Runzhong Wang, Baoyi He, Huaiyong Dou, Junchi Yan, Yongquan Zhang, and Fei Wu. 2024b. Llmco4mr: Llms-aided neural combinatorial optimization for ancient manuscript restoration from fragments with case studies on dunhuang. In *European Conference on Computer Vision*, pages 253–269. Springer.

A Sensitivity Analysis and Intervention-Based Validation of AC-TCGN

We conduct a sensitivity analysis of the neuron selection ratio (ρ) and the activation scaling factor using Qwen3-VL-8B. As shown in Table 6, increasing ρ from 0.03 to 0.05 causes a sharp performance drop, while further increasing it to 0.07 yields marginal effects, indicating that $\rho = 0.05$ is a saturation point that captures the core task-specific subnetwork without excessive redundancy. Figure 5 shows that aggressive inhibition leads to catastrophic degradation on OCR and IC and induces unstable, non-monotonic fluctuations in IU, suggestive of excessive intervention and generation artifacts, whereas overly weak inhibition results in rapid performance recovery, limiting discriminative power. Based on this trade-off, we set the activation scaling factor to 0.1 for all experiments.

Zero-ablation experiments in Table 7 on Qwen3-VL-8B and XunziQwen2-VL-7B produce performance drops comparable to those observed under soft inhibition, further supporting the functional importance of the identified neurons rather than suggesting a purely correlational relationship with task performance. While higher-level semantic tasks such as IU may exhibit absolute performance fluctuations under extremely strong interventions, the relative cross-task interference patterns remain consistent across different neuron selection ratios and scaling factors. These results suggest that the observed entanglement phenomenon is not an artifact of a specific intervention strength but reflects a stable property of the ancient Chinese domain-specific model.

Ratio (ρ)	OCR Acc. (%)	IC Acc. (%)	IU ROUGE-L (%)
0.01	23.88	75.60	6.33
0.03	18.00	38.40	11.01
0.05	2.98	29.60	3.26
0.07	3.32	17.20	3.76

Table 6: Effect of the neuron selection ratio (ρ) on post-inhibition task performance.

Model	OCR (%)	IC (%)	IU (%)	SI
Qwen3-VL-8B	$2.23_{(-85.71)}$	$16.40_{(-61.60)}$	$2.03_{(-6.03)}$	1.81
XunziQwen2-VL-7B	$8.36_{(-91.43)}$	$33.20_{(-14.00)}$	$8.73_{(-14.51)}$	1.29

Table 7: Performance degradation under zero ablation of AC-TCGN-identified neurons on Qwen3-VL-8B and XunziQwen2-VL-7B. SI denotes the separation index between self-task and cross-task interference.

B Neuron Visualization of Qwen2-VL-7B

Figure 6 presents the layer-wise distribution of task-specific neurons for Qwen2-VL-7B under the same experimental settings as in the main experiments. The visualization exhibits qualitative patterns consistent with those observed in Section 6.3 for the general model, including task-dependent layer preferences and partial overlaps across tasks. This consistency suggests that the neuron organization trends reported in the main analysis are not primarily driven by differences in model generation.

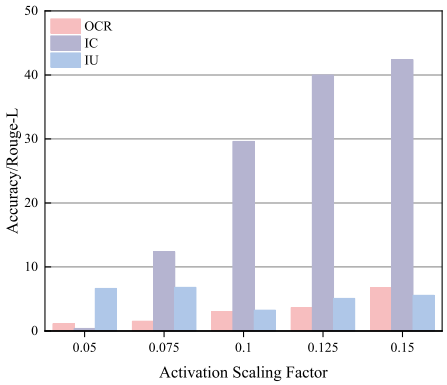


Figure 5: Sensitivity analysis of the activation scaling factor for inhibition of ancient Chinese document task neurons.

C Cross-Task Similarity in the Visual Encoder

Table 8 shows that the visual encoders in both models exhibit high cross-task similarity, indicating

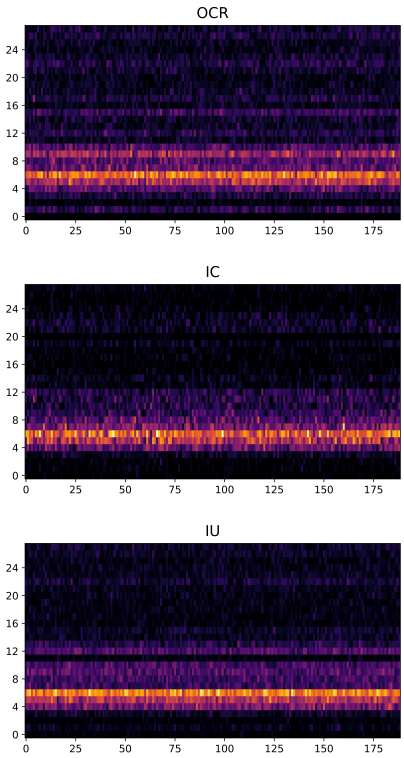


Figure 6: Neuron visualization of Qwen2-VL-7B across three tasks: OCR, IC, and IU.

largely shared and task-agnostic perceptual features. In contrast, decoder-side neuron overlap is substantially lower, suggesting stronger task specificity after visual features are integrated with linguistic context. The observed cross-task coupling therefore appears to arise mainly during decoder-side semantic transformation, where visual forms are mapped onto task-specific meanings. The examined ancient Chinese domain-specific model shows stronger decoder-side coupling than the general model, while visual encoder representations remain relatively similar across tasks. These findings support our dual-chasm hypothesis by showing that perceptual features remain shared, whereas task-dependent divergence and cross-task coupling emerge primarily at the comprehension stage.

Task Pair	Qwen3-VL-8B	XunziQwen2-VL-7B
OCR–IC	0.72	0.75
OCR–IU	0.68	0.71
IC–IU	0.81	0.83

Table 8: Cross-task representational similarity of visual encoder features measured by centered kernel alignment (CKA). Higher values indicate stronger similarity between task representations.

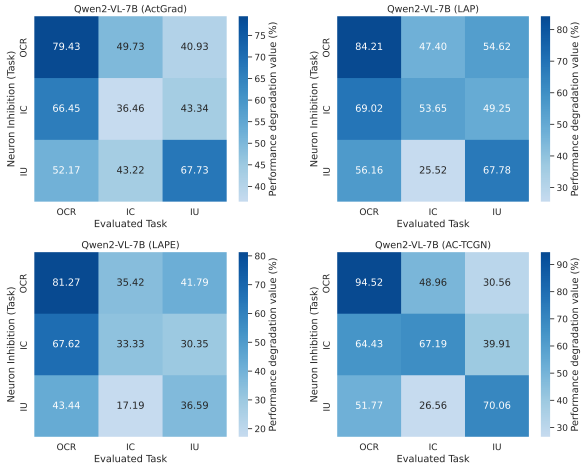


Figure 7: Cross-task effect heatmaps of different neuron identification methods.

D Evaluation of General Language Capabilities

To assess whether the proposed neuron intervention methods primarily target representations specific to ancient Chinese document image–text tasks, rather than impairing the model's general language modeling ability, we evaluate the post-intervention perplexity (PPL) on the Chinese Alpaca dataset. As shown in Figure 8, the base model achieves a PPL of 10.92. Inhibiting OCR-related neurons and IU neurons increases the PPL to 14.58 and 13.39, respectively, while inhibiting IC–related neurons results in the smallest PPL change. Overall, the PPL variations across different neuron inhibition strategies are relatively limited and remain close to the baseline. These results indicate that the intervention methods mainly affect task-specific representation pathways, while largely preserving the general language modeling capability of the model, thereby supporting the targeted design and robustness of the proposed methods.

E Activation Patching Validation

To further assess the functional relevance of neurons ranked by AC-TCGN, we perform clean-to-corrupted activation patching on the held-out validation set. For each example, we corrupt the visual-token representations while keeping the prompt and teacher-forced target sequence fixed, and then restore only the clean post-SwiGLU activations of the top 5% ranked neurons at the corresponding layers and token positions. As shown in Table 9, activation patching produces positive recovery across all six model–task settings, with five of the six settings recovering 66.01%–94.80% of

Model	Task	Clean LL	Corrupt LL	Self-patched LL	Lost LL	Self Δ (%)	Self Recovery (%)
Qwen3-VL-8B	OCR	-0.3440	-1.1854	-0.3877	0.8414	79.77	94.80
Qwen3-VL-8B	IC	-3.2267	-3.9472	-3.3607	0.7205	58.65	81.40
Qwen3-VL-8B	IU	-3.8455	-4.5526	-3.9979	0.7071	55.47	78.45
XunziQwen2-VL-7B	OCR	-0.0635	-0.9130	-0.2008	0.8495	71.22	83.84
XunziQwen2-VL-7B	IC	-3.4773	-4.3446	-4.1559	0.8673	18.87	21.76
XunziQwen2-VL-7B	IU	-2.9704	-3.6956	-3.2169	0.7251	47.87	66.01

Table 9: Clean-to-corrupted activation patching results for AC-TCGN-ranked neurons. LL values are token-normalized within each model–task setting. Lost LL is the corruption-induced loss, Self Δ is the likelihood recovered by patching, and Self Recovery denotes the recovered fraction of Lost LL.

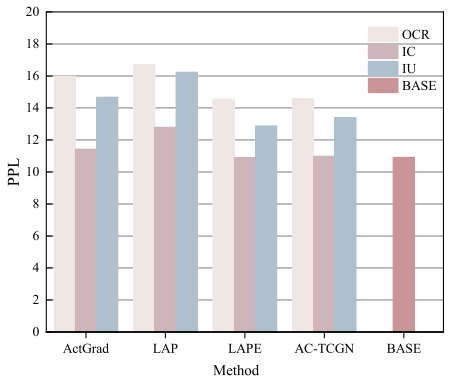


Figure 8: Comparison of PPL variations for ancient Chinese document image–text task neurons under different inhibition methods

the corruption-induced loss in target log-likelihood. The lower recovery on IC for XunziQwen2-VL-7B is consistent with the more distributed IC-related FFN activations observed in Section 6.3. Together with the results of soft inhibition and zero ablation, these activation-patching results provide complementary restoration-based evidence that AC-TCGN identifies neurons whose functional relevance is validated through intervention.