VLHSA: Vision-Language Hierarchical Semantic Alignment for Jigsaw Puzzle Solving with Eroded Gaps (Student Abstract)

Xinyan Liu[1*], **Zhuoning Xu**[1*]
[1]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom, Kowloon-999077, Hong Kong SAR
rowena.liu@connect.polyu.hk, zhuoning-johnny.xu@connect.polyu.hk

Abstract

Jigsaw puzzle solving remains difficult because models must reconcile local fragment cues with global structure. Most prior work leans solely on visual signals (edge or texture coherence) and rarely exploits natural-language descriptions, which are especially helpful for puzzles with eroded gaps. We introduce a vision–language framework that uses textual context to guide assembly. At its core, the Vision–Language Hierarchical Semantic Alignment (VLHSA) module aligns image patches with text via multi-level matching—from local tokens to global summaries—within a multimodal design that couples dual visual encoders with language features for cross-modal reasoning. Across multiple datasets, the method surpasses the state of the art, including a 14.2 percentage-point gain in piece accuracy; ablations identify VLHSA as the principal source of improvement. These results suggest a practical shift for jigsaw solving: augmenting vision with language to resolve ambiguous placements.

Introduction and Previous Work

Solving jigsaw puzzles with eroded gaps demands precise local compatibility and dependable scene level semantics. Purely visual pipelines are often confused by worn edges and visually similar fragments. Pairwise relation learning, including Deepzzle (Paumard, Picard, and Tabia 2020) which uses a Siamese network (Bertinetto et al. 2016) to classify the relative position of neighbors around a center piece, captures local fit but rarely enforces global coherence, especially when fragments are ambiguous or eroded. Reinforcement learning planners such as SD^{2}RL (Song et al. 2023a) improve end to end placement but struggle to scale to large puzzles. Discriminator guided search exemplified by PDN-GA (Song et al. 2023b) is robust yet computationally heavy and more brittle under large gaps. Multiscale structural modeling such as ERL-MPP (Song et al. 2025) gathers local to global cues, but vision only reasoning remains ambiguous on similar pieces. Generative reconstruction like JPDVT (Liu et al. 2024) conditions diffusion and ViT models to tolerate missing pieces, with trade offs in efficiency and stability when erosion is severe. We present *VLHSA*, a hierarchical alignment module for vision and language that links fragments to textual cues from local tags up to global descriptions so semantic priors guide placement. On standard and gap eroded benchmarks, VLHSA yields consistent gains, up to 14.2 percentage points gain in piece accuracy, and ablations indicate that hierarchical semantic guidance is the main driver.

[*]These authors contributed equally.

Proposed Methodology

Problem Formulation

We address jigsaw reconstruction with eroded gaps, where an image is partitioned into N fragments and the goal is to recover a one-to-one piece–to–grid permutation. Beyond low-level visual cues, a concise caption $\mathcal{T}$ provides semantic constraints that help disambiguate visually similar fragments and enforce global coherence.

Framework Architecture

Our framework couples a dual visual encoder with a lightweight vision–language alignment module. Vision Mamba captures long-range spatial dependencies, while a BLIP vision encoder supplies language-aligned semantics; features are projected to a common space and combined via residual adapters to remain parameter-efficient. On the text side, a short caption is generated using BLIP's captioning capability and encoded into token-level and global embeddings (CLIP text encoder), resulting in a unified multimodal representation with minimal overhead.

Vision–Language Hierarchical Semantic Alignment

The core VLHSA module aligns visual fragments and language at three granularities. (i) *Token level*: cross-attention links patches to salient words, enabling fragments to focus on relevant concepts. (ii) *Region level*: pooled neighborhoods (e.g., $2\times2/3\times3$) are matched to phrase spans, capturing local context and spatial relations. (iii) *Global level*: a gating mechanism reconciles image-wide semantics with the full caption to ensure holistic consistency. Aligned features are fused with learnable weights to balance fine-to-global signals adaptively.

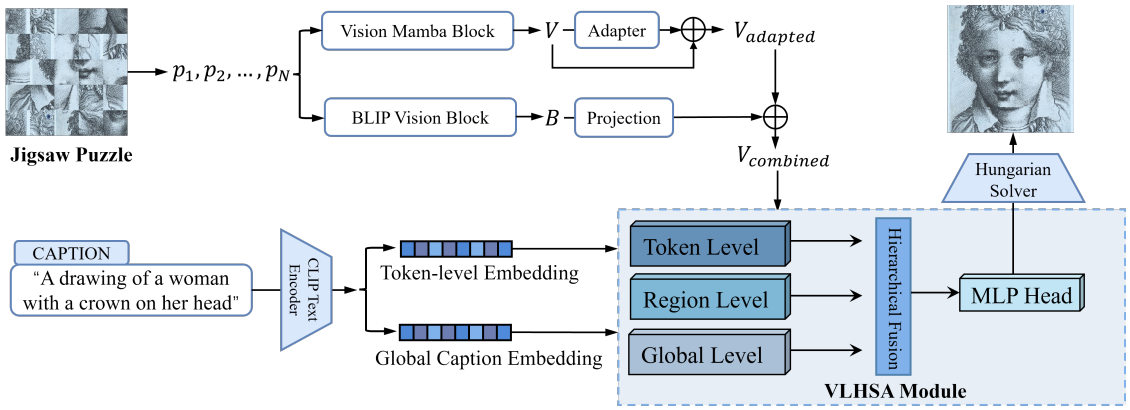

Figure 1: VLHSA overview: Mamba + BLIP features, CLIP text (global/tokens), hierarchical alignment (token/region/global), Hungarian assignment.

Assignment and Training Objective

Fused representations are fed to a lightweight prediction head that outputs position scores; a Hungarian matching layer enforces bijective assignment. The training objective combines assignment accuracy with mild regularizers that stabilize the three alignment stages and a small pairwise adjacency term for local consistency. At inference, only a brief caption is required; the model performs a single forward pass to produce the final permutation.

Our total training objective combines reconstruction accuracy with semantic alignment, and includes minor auxiliary terms for pairwise adjacency and local consistency regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{assign}} + \lambda(\mathcal{L}_{\text{token}} + \mathcal{L}_{\text{region}} + \mathcal{L}_{\text{global}}) + \lambda_{\text{p}}\mathcal{L}_{\text{p}} \quad (1)$$

where $\mathcal{L}_{\text{assign}}$ is the cross-entropy loss for the optimal assignment, and λ balances reconstruction accuracy with semantic coherence. For regularization, we introduce a pairwise adjacency loss ($\mathcal{L}_{\text{p}}$), which serves as minor auxiliary terms without affecting the main vision-language alignment framework.

Experiments

Datasets and Protocol

We evaluate on JPwLEG-3 and JPwLEG-5, two jigsaw benchmarks with eroded gaps derived from the MET open-access collection. JPwLEG-3 uses 3×3 fragments; JPwLEG-5 uses 5×5 with larger gaps and higher ambiguity. We follow the official splits and the training/evaluation protocol introduced by SD^{2}RL, ensuring fair comparison to prior work. Metrics include Perfect, Piece, Horizontal, and Vertical accuracy. Further data preprocessing details and experimental settings are deferred to the appendix.

Main Results

On the harder JPwLEG-5, VLHSA establishes a new state of the art in *Perfect* accuracy (19.0%) while delivering a substantial gain in *Piece* accuracy (66.9%), a 14.2 pp improvement over ERL-MPP (52.7%). Horizontal/Vertical relations also improve to 58.1%/58.4%. On JPwLEG-3, VLHSA attains 85.4% *Piece* accuracy, surpassing SD^{2}RL (81.6%). Category-wise analysis shows the largest gains on engravings and artifacts (15.0 pp over the strongest baseline), highlighting stronger robustness under fine-grained structure and high visual similarity.

Method	Venue	Perf.	Piece	Hori.	Vert.
Deepzzle	TIP-2020	0.0	21.9	10.9	10.7
SD^{2}RL	AAAI-2023	5.1	40.3	26.5	26.2
PDN-GA	ICASSP-2023	6.1	44.3	30.8	30.6
ERL-MPP	AAAI-2025	18.6	52.7	56.5	57.3
VLHSA (Ours)	-	**19.0**	**66.9**	**58.1**	**58.4**

Table 1: JPwLEG-5 comparison (accuracy %). Full baselines and per-category results are in the appendix.

Encoder ablations show progressive gains when augmenting Vision Mamba with BLIP visual features and CLIP text, confirming the utility of cross-modal cues. The best results are obtained when all three levels are enabled. Complete tables are provided in the appendix.

References

Bertinetto, L.; Valmadre, J.; Henriques, J. F.; Vedaldi, A.; and Torr, P. H. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, 850–865. Springer.

Liu, J.; Teshome, W.; Ghimire, S.; Sznaier, M.; and Camps, O. 2024. Solving Masked Jigsaw Puzzles with Diffusion Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23009–23018.

Paumard, M.-M.; Picard, D.; and Tabia, H. 2020. Deepzzle: Solving Visual Jigsaw Puzzles With Deep Learning and Shortest Path Optimization. *IEEE Transactions on Image Processing*, 29: 3569–3581.

Song, X.; Jin, J.; Yao, C.; Wang, S.; Ren, J.; and Bai, R.

2023a. Siamese-Discriminant Deep Reinforcement Learning for Solving Jigsaw Puzzles with Large Eroded Gaps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2): 2303–2311.

Song, X.; Yang, X.; Ren, J.; Bai, R.; and Jiang, X. 2023b. Solving Jigsaw Puzzle of Large Eroded Gaps Using Puzzlet Discriminant Network. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Song, X.; Yang, X.; Yao, C.; Ren, J.; Bai, R.; Chen, X.; and Jiang, X. 2025. ERL-MPP: Evolutionary Reinforcement Learning with Multi-head Puzzle Perception for Solving Large-scale Jigsaw Puzzles of Eroded Gaps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7): 6968–6977.