LLMCO4MR: LLMs-aided Neural Combinatorial Optimization for Ancient Manuscript Restoration from Fragments with Case Studies on Dunhuang

Yuqing Zhang[1], Hangqi Li[1], Shengyu Zhang[1,2]*, Runzhong Wang[3], Baoyi He[1], Huaiyong Dou[1], Junchi Yan[3,4]*, Yongquan Zhang[1], and Fei Wu[1,2,4]

[1] Zhejiang University, Hangzhou, China
[2] Shanghai Institute for Advanced Study of Zhejiang University, Shanghai, China
[3] Shanghai Jiao Tong University, Shanghai, China
[4] Shanghai AI Laboratory, Shanghai, China
{yuqingz7, sy_zhang, wufei}@zju.edu.cn, yanjunchi@sjtu.edu.cn

Abstract. Restoring ancient manuscripts fragments, *such as those from Dunhuang*, is crucial for preserving human historical culture. However, their worldwide dispersal and the shifts in cultural and historical contexts pose significant restoration challenges. Traditional archaeological efforts primarily focus on manually piecing major fragments together, yet the small and more intricate pieces remain largely unexplored, which is technically due to their irregular shapes, sparse textual content, and extensive combinatorial space for reassembly. In this paper, we formalize the task of restoring the ancient manuscript from fragments as a cardinality-constrained combinatorial optimization problem, and propose a framework named LLMCO4MR: (Multimodal) Large Language Model-aided Combinatorial Optimization Neural Networks for Ancient Manuscript Restoration. Specifically, LLMCO4MR encapsulates a neural combinatorial solver equipped with a differentiable optimal transport (OT) layer, to efficiently predict the Top-K likely mutual reassembly candidates. Multimodal Large Language Model (MLLM) is then adopted and prompted to yield pairwise matching confidence and relative directions for final restoration. Experiments on synthetic data and cases studies with real-world famous Dunhuang fragments demonstrate our approach's practical potential in assisting archaeologists. Our method provides a novel perspective for ancient manuscript restoration.

Keywords: Ancient Manuscript Restoration · Dunhuang Manuscript · Combinatorial Optimization · Multimodal Large Language Models

1 Introduction

The restoration of ancient manuscripts has long been the focus of archaeologists' research and the global community. Traditionally, these codices are frequently

* Corresponding Authors.

(a) Conventional manuscript restoration task (b) **Our task:** Small fragments restoration with outliers

Fig. 1: Comparison of conventional manuscript restoration and our task. (a) highlights the conventional restoration task with large fragments. Archaeologists focus on written content as primary guide for restoration, supplemented by texture and edges. (b) outlines our task. In an open-world problem setting, small fragments contain limited text. Recovery relies on contour similarity matching. This is generally difficult for humans to identify fragment pairs from a large collection containing outliers.

discovered in a (severely) deteriorated condition due to the passage of time. Recovering these manuscripts involves extensive efforts [37] that are often time-consuming, labor-intensive, and require a high level of expertise, necessitating years of professional training [2].

On the other hand, vision technologies have demonstrated significant potential in aiding the restoration of ancient manuscripts. For example, [1] developed a sophisticated system based on Convolutional Neural Network (CNN) to identify matching candidates among papyrus fragments by analyzing their fiber patterns. Zhang et al. [51] proposed a self-supervised model using Generative Adversarial Network (GAN) and siamese network for the pairwise stitching of Oracle Bone pieces, utilizing manually annotated contours for similarity comparison. While these vision-assisted pipelines enhance efficiency and accuracy beyond traditional manual efforts, their applications are generally confined to local pairwise matching [31], or regularly-shaped (e.g., rectangle) and large fragments [54]. A major challenge persists in dealing with real-world small fragments, which are often less-textual, shape-irregular, and scattered with missing fragments.

These small fragments have often been overlooked by archaeologists, yet they form a substantial part of the manuscript collections. Recent explorations have revealed that these minor pieces are treasure troves of first-hand materials, offering unique insights into the daily lives of past civilizations [12]. Current deep learning approaches rely on textual [29], contextual [2] or multimodal information [54] to reassembly manuscripts. Such learning-based methods require a large amount of annotated training data for complex multimodal feature extraction, which is impractical for small fragment restoration.

To bridge the gap, we resort to the recent advancements in large language models (LLMs) that have revolutionized traditional tasks in computer vision and natural language processing [7, 26, 38]. Notably, the emergence of Multimodal Large Language Models (MLLMs) has demonstrated exceptional capabilities in extracting and interpreting image features, understanding content, and following complex prompt instructions [4, 20, 48]. Capitalizing on these developments, our

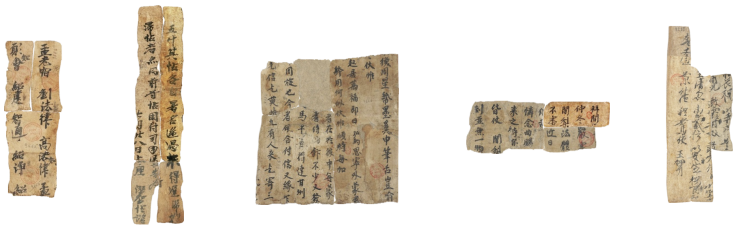

Fig. 2: Showcases of matching results with case studies on Dunhuang fragments.

research introduces a novel approach by employing MLLMs to tackle the intricate task of restoring ancient manuscript fragments. MLLMs facilitate efficient pairwise matching by: 1) identifying cross-image connections based on contour similarity and capturing complex features similarity in paper texture and content; 2) replicating the comprehensive integration of multimodal semantics that archaeologists employ during their assembly process; 3) demonstrating zero-shot learning capabilities, which are crucial for investigating manuscripts with limited ground truth availability.

However, the complexity of our task surpasses the current capabilities of MLLMs. In practical scenarios of ancient manuscript restoration, it is necessary to first identify potential candidate fragments from a larger collection for subsequent pairwise matching. This collection contains fragments that may not belong to the target manuscript, which we refer to as *"outliers"*. Performing pairwise comparisons using MLLMs on the vast number of small fragments, including these outliers, leads to a ***combinatorial explosion*** challenge, which poses a significant difficulty for MLLM itself to manage effectively.

To address these challenges, we formulate the fragment selection subtask as a combinatorial optimization (CO) problem. We introduce a novel framework, named LLMCO4MR, a LLMs-aided Neural Combinatorial Optimization solver for manuscript restoration, transforming the unstructured challenge into a technically feasible task. **The highlights of the paper are:**

1. **More Open-world Problem Setting.** Beyond the restricted setting requiring that there are no outlier fragments for recovery, to our best knowledge, we are the first to address the more realistic and challenging setting, i.e., restoring the manuscript from a subset of fragments with outliers.
2. **New Two-stage Solving Pipeline.** We solve the new problem by two stages: fragment selection and pairwise matching. In the first stage, k fragments with the highest mutual assemble scores are selected by certain means while the latter stage involves the traditional closed-world problem setting.
3. **Neural Combinatorial Solver for Candidate Selection.** We introduce a combinatorial network to solve the fragment selection problem by formulating it as a K-cardinality constrained task, with an advantage that the input visual fragments are learned end-to-end directly with the final decision goal. Supervised training is achieved by using the shredded fragments generated by splitting a complete manuscript into grids.

4. **MLLMs for Pairwise Fragment Matching in Zero-shot.** The selected Top-K fragments are then fed into the MLLM by pair, to generate the matching score as well as the relative direction of matching, i.e., horizontal or vertical. Then the final manuscript can be recovered by these local pairwise matchings. To our best knowledge, this is the first time to introduce MLLMs for manuscript recovery or more broadly speaking, jigsaw puzzle-like tasks.

2 Related Works

Manuscript Restoration. This area has recently witnessed significant advancements [2, 37] aided with vision technology for restoration. These developments aim to overcome the challenges in manual methods, which are time-consuming, labor-intensive, and heavily reliant on expert knowledge [2]. A typical method is to solve the problem from a jigsaw puzzle [10, 34] solving perspective. For puzzle solving and global restoration, several successful methods have emerged. Traditional strategies include curve-based techniques [27, 44] and time series sequence-based approaches [6, 13, 32, 39, 53]. Additionally, deep learning methods have been developed, exemplified by works such as [19] and other studies [17, 25, 28]. Inspired by these works, previous manuscript restoration methods are designed based on the visual features of the fragments. Here, *matching* often refers to identifying whether two fragments are adjacent and their relative direction (horizontally or vertically). In particular, the feature extraction module has been further advanced by deep learning with various aspects, including fragment contour shape [51, 52, 55], texture details [31], and written characters [29]. In the recent work [54], a multimodal restoration pipeline is devised that considers both written content and contour shape.

Despite these advances, most current relic restoration methods share a common assumption: the candidate pieces for pairwise matching are given, either by manual pre-selection or other means. However, in many real-world restoration cases, this assumption leaves a gap in addressing the more challenging task of open-world problem setting, where candidate fragments are not predetermined, underscoring the novelty and necessity of our approach.

Combinatorial Solver. The recent growing trend of learnable CO solvers [5] demonstrate the potential of solving CO problems by deep neural networks with higher efficiency [18, 22, 40, 42]. To tackle the extensive combinatorial space of ancient manuscript restoration, we present a cardinality-constrained CO reformulation [8] by selecting the Top-K most likely matching fragments from the candidate set. Following the one-shot CO learning paradigm [41], we adopt [43], which handles cardinality-constrained problems similarly to differentiable optimal transport [46]. [43] has the advantage of controlling the tightness of constraints while preserving the gradient, which is crucial in our task.

Large Language Models. With the recent arisen success of Large Language Model (LLM), such as GPT-4 [26], Qwen [3] and Baichuan [47], LLMs have showcased their efficacy in a myriad of fields including computer vision [23, 24], natural

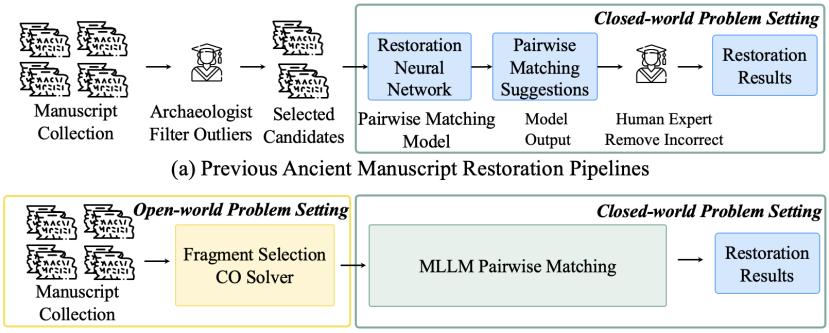

Fig. 3: Comparison between previous restoration approaches and our LLMCO4MR.

Table 1: Comparison with various ancient fragment restoration SOTA.

Methodology	Training Data	Annotation Free	Eliminate Outliers	Network Structure	Extracted Features
JigsawNet [19]	Self-constructed	✓	✗	CNN	Texture
Papyrus [31]	Self-constructed	✗	✗	Siamese CNN	Texture
S3-Net [51]	GAN-augmented	✗	✗	Siamese CNN	Contour
LLMCO4MR	Synthetic Shredding	✓	✓	CO-MLLM	Multimodal

language processing [38], medical application [35,36], law judgement [33,50], and optimization [15,16]. Recent advancements in Multimodal-LLM (MLLM) have catalyzed a proliferation of diverse applications, including image content understanding [56], mathematical reasoning [11], and even meme interpolation [49], among others. Inspired by MLLMs' potential in complex visual and textual content understanding, we proposed LLMCO4MR, the first CO solver and MLLM collaboration pipeline to tackle the complex ancient manuscript restoration task. We leverage MLLMs' potential extracting multimodal features [14,24] to tackle small ancient fragments pairwise matching, while further enhancing CO solver performance figuring similar fragments in real-world cases.

Discussion with Other Restoration Pipelines. We provide a qualitative comparison with other major jigsaw puzzle solving and fragment restoration pipelines, as shown in Tab. 1. The key distinctions between our LLMCO4MR pipeline and earlier approaches are highlighted in Fig. 3. Our pipeline adeptly addresses the combinatorial explosion challenge inherent in fragment selection through CO solver. Furthermore, it innovates by introducing the use of MLLMs for the pairwise matching of ancient manuscripts, an innovative approach for manuscript restoration with open-world setting.

3 Methodology

In this section, we provide a detailed description of our LLMCO4MR pipeline, as composed of two main components: a cardinality-constrained neural solver trained on synthetic data to select candidates for further matching; and the MLLM adoption to achieve effective matching in a zero-shot learning manner.

3.1 Problem Formulation

We consider a set of m fragments, denoted as $\mathcal{F} = \{f_{in}^1, \ldots, f_{in}^k, f_{out}^{k+1}, \ldots, f_{out}^m\}$. Here the first k fragments f_{in}^i refer to those originating from a more complete fragment, while the rest $(m-k)$ fragments denotes those outliers. The goal of LLMCO4MR is to identify such k relevant fragments.

3.2 CO Solver for Fragment Selection

3.2.1 Preliminaries. Based on consultations with archaeology specialists and their insights gained from experimental endeavors, we have formulated the following hypotheses to guide the restoration:

1. **Internal Texture and Limited Text.** Small real-world fragments typically contain very limited characters, and ancient text recognition is unreliable for matching due to low accuracy. Moreover, even fragments that should be joined together may exhibit different textures due to varying museum collection conditions. Therefore, distinguishing fragments by textual content or texture patterns poses a significant challenge.
2. **Edge Shapes and Historical Tearing.** Small fragments were often deliberately torn by people in ancient times [12], resulting in contours that are not severely eroded. This key feature offers the potential for primary restoration through contour similarity.
3. **Reassembly Group Size.** Fragments are typically reassembled in groups ranging from 3 to 9 pieces. It is uncommon to find a large assembly of pieces that can be combined into a significantly larger fragment.

In this section, we propose a cardinality-constrained neural CO solver to efficiently and robustly select manuscript fragments, which is further exemplified with MLLM with application on the famous real-world Dunhuang Manuscript.

3.2.2 Solver Overview. We formulate fragment selection task as cardinality-constrained optimization problem in Eq. (1a). Specifically, the number of non-negative elements must be restricted to no greater than k, while minimizing the objective L_{co}. The aim of our formulation is to find the optimal fragments subset with a size of k with highest mutual similarity, as shown in Eq. (1b), where D_{ij}

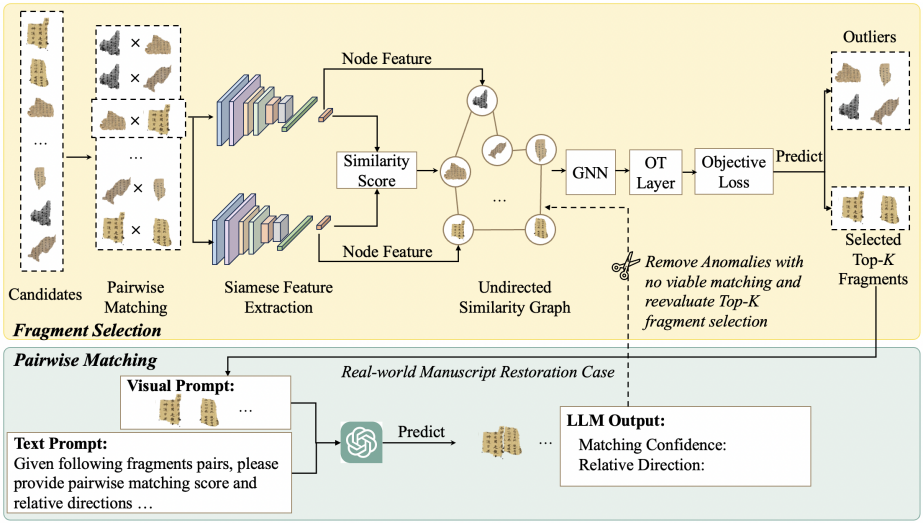

Fig. 4: LLMCO4MR solving pipeline. 1. Fragment Selection with CO solver. Given a set of m fragments, CO solver first pairs and feeds them into a siamese network to obtain similarity. Next, construct similarity graph $\mathcal{G}$ and pass it through GNN. Then encode cardinality constraints with differentiate OT layer. Finally output the selected Top-K fragments with highest mutual similarity. **2. Pairwise matching with MLLM in zero-shot.** Given the selected Top-K fragments by pair as visual prompt with text prompt, MLLMs generate matching score and relative directions as output.

denotes the similarity score between candidates i and j, which are predicted by siamese network as described in the following section.

$$\min_{\mathcal{S}} \; L_{co} \quad \text{s.t. } \mathcal{S} \in \{0,1\}^m, \quad \|\mathcal{S}\|_0 \leq k, \tag{1a}$$

$$\text{where } L_{co} = -\sum_{i \in \{i|s_i=1\}} \left(\sum_{j \in \{j|s_j=1, j \neq i\}} D_{ij} \right). \tag{1b}$$

This cardinaly-constrained CO problem could be tackled by a differentiable OT layer [43] and Sinkhorn algorithm [9]. As shown in Fig. 4, we construct an end-to-end learning pipeline consisting of a siamese feature extraction module as problem encoder network, a graph neural network fused with an OT layer as the solver network. The neural network learns to solve Eq. (1a) under the unsupervised loss described in Eq. (3) as follows.

3.2.3 Siamese Feature Extraction Network. In this section, we introduce the siamese feature extraction network of CO solver. Given the scarcity of real-world ground truth data, as few archaeologists have successfully restored small manuscript fragments together, we first describe the synthetic shredding process for preparing training data.

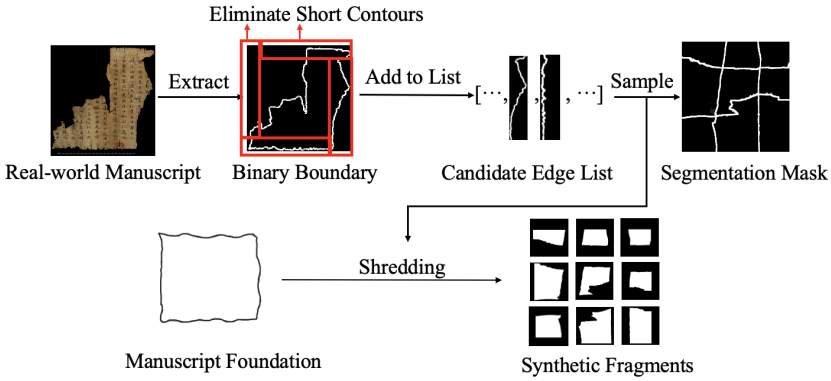

Fig. 5: Illustration of synthesized shredding process.

Synthetic Shredding Process. We draw an analogy between the intricate work of manuscript restoration and grid puzzle assembly. To simulate this, we generate synthetic shredded fragments by breaking larger manuscript pieces into smaller, puzzle-piece-like fragments. Our synthesis approach is conducted in two phases, as illustrated in Fig. 5:

1. **Contour Extraction.** We begin by randomly selecting large, real-world manuscripts from which we extract binary boundaries, specifically focusing on the two horizontal and two vertical edges. We discard contours that are either too long or too short, retaining only those of an appropriate length to form our contour set $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$. Each contour in this set is represented by a series of two-dimensional coordinates $C_i = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Through this method, we ensure the synthetic fragments are realistic in size and shape.
2. **Synthetic Shredding.** We proceed to sample from the set $\mathcal{C}$ without replacement to create an $M \times M$ grid-like segmentation mask. Utilizing a real rectangular-shaped manuscript piece as our foundation, we apply this mask to segment it into M^2 smaller, discrete fragments. This division simulates realistic fragment sizes and shapes encountered in manuscript restoration.

Siamese Network. We employ a siamese feature extraction network designed to assess contour similarity. We start with the original set of fragments with outliers $\mathcal{F} = \{f_{in}^1, \ldots, f_{in}^k, f_{out}^{k+1}, \ldots, f_{out}^m\}$. For each fragment, we create permutation combinations of pairwise matches (f^i, f^j). These paired fragments, specifically their masked contour, are then fed into two identical branches of a CNN-based neural network. The network computes a similarity score $D_{ij} = Siamese(f^i, f^j)$ with the range being [0-1]. A score of 1 indicates that two fragments have matching contours that align perfectly, suggesting adjacency. Conversely, a score of 0 implies not adjacent. Eq. (2) describes the loss term, w means learnable parameters, $L_{siamese}$ means Binary Cross Entropy (BCE) Loss,

p means predicted similarity score and y means ground truth label.

$$L_{siamese} = -w\left[y \cdot \log p + (1-y) \cdot \log\left(1-p\right)\right]. \tag{2}$$

3.2.4 Solver Network. Our CO solver network consists of a graph neural network fused with a differentiable OT layer.

Graph Neural Network. Following the siamese feature extraction network, we construct an undirected similarity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Graph vertices set $\mathcal{V} = \{v_1, \ldots, v_m\}$ stores the extracted fragment mask feature vector. Given the pairwise similarity scores D_{ij} as edge weights, we add an edge e_{ij} to the edge set $\mathcal{E}$ if similarity score is greater than threshold. We then constructed a graph neural network to extract features from graph vertices and edges. We use Graph Convolution Network (GCN) as backbone.

OT Layer and training. For Top-K fragments selection, we encode the cardinality constraint with differential OT layer following [41]. To encode cardinality constraints in a differentiable manner, we first formulate the constraints as marginal distributions in optimal transport (OT), and then solve OT with Sinkhorn algorithm [9] by iterative row-wise and column-wise normalization. Given m GNN output score p_i indicating the probability of selecting each piece, OT layer moves Top-K probability to p_{max} while moves others to p_{min}. We select K nodes with the highest probability to predict the output and optimize neural network with loss function in Eq. (3). Loss function consists of three components: $L_{siamese}$ penalize incorrect prediction in pairwise similarity; L_{gnn} denotes the BCE loss for node selection; and L_{co} indicates the optimization objective in the previously mentioned cardinality constrained CO.

$$\min L = \min(\alpha L_{siamese} + \beta L_{gnn} + L_{co}). \tag{3}$$

3.2.5 Second-order Matching As detailed in Section 3.1, fragment groups ranging from 3 to 9 pieces can be reconstructed into a final form. We explore different scenarios of grid-like restoration. In the case of a 2×2 grid, each ground truth fragment is expected to have two pairwise connections with its neighboring fragments, treating all four fragments as equally important. In contrast, within a 3×3 puzzle grid, as illustrated in Fig. 6, fragments positioned at [2, 4, 6, and 8] are observed to share three connections each, highlighting the central fragment [5] as most crucial with up to four connections. Meanwhile, the corner fragments [1, 3, 7, 9] are limited to just two connections. Acknowledging the imperfection of siamese network accuracy and the existence of false positive pairwise matches in real-world scenarios, we implement a *second-order matching strategy*. We start with identifying the Top-5 central pieces that with a higher number of connections. Subsequently, we refine the graph $\mathcal{G}prune$ by pruning edges, retaining only those connected to the previously selected Top-5 fragments and eliminating other distracting edges. This strategy is designed to mitigate the impact of inaccuracies and ensure a more reliable restoration process.

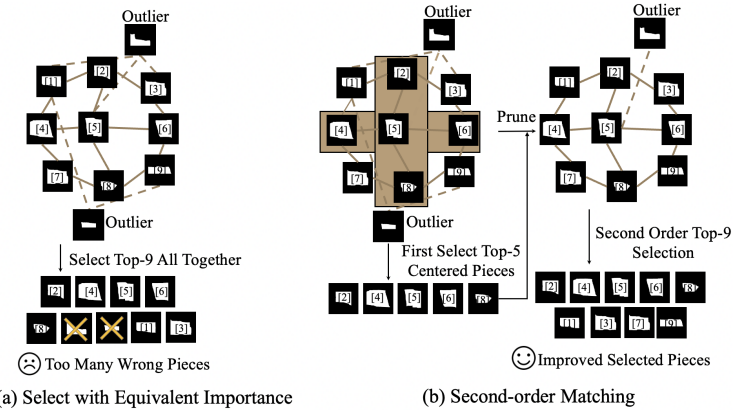

Fig. 6: Illustration of second-order matching. (a) selects all Top-9 fragments with equivalent importance (dashed lines represent incorrect pairings obtained by siamese network). (b) highlights centered 5 fragments with more connections, then select the rest 4 fragments based on pruned graph $\mathcal{G}_{prune}$ (most incorrect pairings are eliminated).

3.3 LLMCO4MR Pipeline for Manuscript Restoration

Traditional deep learning methods for manuscript restoration heavily depend on human intervention to filter out numerous incorrect pairings generated by models. This task becomes increasingly challenging with smaller fragments, which are abundant and tiny, making manual review burdensome.

The advent of MLLMs offers high-quality, efficient solutions to tasks that were previously labor-intensive. Innovatively, we are among the first to harness MLLMs for manuscript restoration. MLLMs like GPT-4V have shown remarkable capabilities akin to those of archaeologists by incorporating multimodal semantics for precise pairwise matching, excelling in identifying local pairwise connections with scoring and directional guidance. However, MLLMs struggle with the vast possibilities in open-world settings, particularly with outliers, limiting their ability to perform restoration independently.

Addressing these challenges, we present the LLMCO4MR pipeline, a pioneering MLLM-aided Neural Combinatorial Optimization solver tailored for the restoration of small manuscript fragments. LLMCO4MR leverages a CO solver for initial fragment selection, effectively countering the combinatorial explosion by sidelining outliers. Subsequently, the Top-K selected pairs undergo analysis via MLLMs, which utilize custom prompts based on a role-play strategy [45] to refine matching precision. If MLLMs identify any Top-K selected fragments as having no viable pairing, these anomalies are removed from the CO solver's input, suggesting a reevaluation of the Top-K fragment selections. This iterative refinement enhances LLMCO4MR efficacy. Experimental results in Section 4.3, demonstrate how this synergistic approach not only improves the precision in selecting fragments but also significantly enhances the accuracy and efficiency of manuscript restoration in practical scenarios.

4 Quantitative Evaluation and Case Study

Our experiments encompasses both synthetic data and real-world manuscript fragments. For fragment selection, we conducted **simulation experiments** using the CO solver, to demonstrate its capability in selecting appropriate fragment candidates for subsequent pairwise matching. We then deployed our LLMCO4MR pipeline on real-world scenarios. Specifically, we conducted **case studies** on the famous Dunhuang manuscripts, spanning from the 4th to the 14th century. These manuscripts are highly valued by archaeologists and the global community [12], due to their extensive volume, diverse content, and broad historical span.

The combination of simulation experiments and real-world case studies provides a controlled and ideal environment for the quantitative evaluation of our method, aligning with the protocols of peer studies [30, 31, 51, 54]. These experiments were performed on a server equipped with Intel i7 CPUs at 3.2GHz and NVIDIA GeForce RTX 3090 GPUs.

4.1 Experiment Setup

Data. For fragment selection, simulation experiments were employed to train the CO solver using synthetic shredded data, as elaborated in Section 3.2.3. Our methodology ensured that fragments generated from actual manuscripts, including their contours, accurately reflect the complexities found in real-world restoration conditions. For the case studies involving LLMCO4MR pipeline, high-quality (with annotated ground truth) real-world Dunhuang fragments are indeed scarce. Our annotated fragment cases were provided by the expert team dedicated to Dunhuang culture research, in the form of visual images for evaluation. All images were obtained from the International Dunhuang Project (IDP).

Evaluation. For fragment selection with CO solver, we evaluated on synthetic shredded data with accuracy criterion. For instance, in a scenario where the task is to select Top-4 pieces from 10 candidates, fragment correctly identified as one of the 4 ground truth fragments is deemed a successful selection. For the real-world case study on Dunhuang, we evaluate the performance by measuring the success rate of fragment restoration. A successful restoration is defined as a match that aligns with the ground truth candidate pairings provided by archaeologists. This expert-annotated data serves as our benchmark for assessing the accuracy of our method in real-world scenarios.

4.2 Simulation Experiment on Combinatorial Solver

Training Details. Aligned with Section 3.1, CO solver is designed to select 2×2 and 3×3 fragments from a set $\mathcal{F}$ comprising m fragments, aiming to address the Top-K fragment selection challenge. Thus we initially selected k pieces that originate from the same fragment and then added $(m - k)$ random fragments as outliers to compile a group. We assembled 500 such groups as

Table 2: Performance of CO solver siamese feature extraction network.

Data Quantity	Hybrid Ratio	Siamese Accuracy	Top-4 Performance
40000	1:1	0.8746	0.6925
	2:3	0.9212	0.7025
	1:2	0.9117	0.7100
80000	1:1	0.9534	0.7250
	2:3	0.9437	0.7325
	1:2	0.9675	0.7500
120000	1:1	0.9245	0.7525
	2:3	0.9586	0.7575
	1:2	**0.9731**	**0.7800**

Table 3: Performance of CO solver on various Top-K fragments selection task.

Top-K	2×2			3×3		
Pool Size	10	15	20	20	25	30
Random Select	0.4000	0.2667	0.2000	0.4500	0.3600	0.3000
Baseline	0.7475	0.3125	0.2575	0.6756	0.5767	0.5389
OT Layer	**0.7800**	0.5825	0.4225	**0.7256**	0.6444	0.6044

training dataset, reserving an additional 100 groups for evaluation. We trained the siamese feature extraction network with 40K, 80K and 120K positive and negative pairing samples, with the ratio of training and testing being 6:1.

Siamese Feature Extraction Network. As detailed in Table 2, we evaluate the performance across various training data quantity, positive and negative sample ratios utilizing a ResNet backbone. Our findings indicate that CO solver performance peaks on Top-4 selections with 120K samples and 1:2 hybrid ratio on positive and negative matching samples. This can be attributed to the imbalance between positive and negative samples in pairwise matching manuscript fragments. With more negative samples, the reliability of siamese model is improved and CO solver performance is enhanced.

CO Solver. Tab. 3 presents the simulation experiment results under various configurations of Top-K fragment selection from m candidates. "Random select" refers to the Top-K random selections mathematical expectation. "Baseline" refers to selecting the Top-K probabilities in a greedy manner from GNN output without OT layer. The introduction of the OT layer, which enforces cardinality constraints, is observed to enhance model performance, particularly as the candidate size m increases. This improvement is attributed to the OT layer's rigorous penalization of deviations from the optimization objectives in Eqs. (1a) and (1b).

Table 4: Enhancement from the second-order Top-K selection.

Strategy	Top-9 of 20	Top-9 of 25	Top-9 of 30
Select at Once	0.7256	0.6444	0.6044
Second-Order	**0.7311**	0.6822	0.6633

Enhancement through Second-order Matching Strategy. In the 2×2 selection settings, all four ground truth fragments hold equal significance (each fragment only have exactly 2 matching connections with their ground-truth adjacent pieces). Thus we explore the efficacy of the second-order matching strategy introduced in Section 3.2.5 for 3×3 scenarios and detail the results in Tab. 4. The second-order matching strategy duplicates the cardinality constraint and applies more stringent penalties for incorrect selections. This method aims to mitigate the effects of erroneous pairwise matches proposed by the siamese network, thereby advancing the accuracy and reliability of fragment selection.

Simulation experiments demonstrate our proposed neural CO solver's effectiveness in selecting candidates based on highest mutual assemble similarity. This gives promise for the subsequent real-world restoration with LLMCO4MR.

4.3 Real-world Manuscript Case Studies on Dunhuang

Results with LLMCO4MR. We conducted case study experiments utilizing LLMCO4MR and compared its effectiveness against other SOTAs, as detailed in Tab. 5. We input a group of fragments with outliers to CO solver for selection, then use the selected Top-K fragments as image prompts for GPT-4V, which suggests the matching scores and relative directions. For SOTAs [31,51] that lack the capability to filter out outliers, we repeatedly input the fragments pairs and select the matching results in a greedy manner (select the fragment candidate pairs with ranked highest matching score).

We observed that previous works often operated under restricted conditions, such as assuming no outliers, limiting search spaces, regular shape fragments, or using manually outlined perfect matching contours. These constraints do not reflect the open-world setting addressed by our method. Moreover, contour similarity network does not achieve perfect accuracy in matching judgments and real-world scenarios frequently involve more irregular fragments, underscoring the complexity of the task.

Our pipeline addresses these challenges in two ways. First, the CO solver captures irregularity through data augmentation. We extract contours from real-world Dunhuang manuscripts and apply to shredding to adequately represent real-world challenges and mitigate the risk of overfitting unrealistic characteristics in training. Second, the MLLM component captures cross-image similarity using multimodal features, including contour (even with some imperfect fits and irregular shapes) and other visual features.

Table 5: Evaluation on real-world Dunhuang fragment restoration case studies.

Top-K	2×2		3×3	
Pool Size	10	15	20	25
GPT-4V [26]	0.3250	0.1750	0.3778	0.2533
LLaVA [21]	0.3150	0.1250	0.2778	0.1556
JigsawNet [19]	0.5250	0.3750	0.4556	0.3222
Papyrus [31]	0.4250	0.3750	0.4778	0.4111
S3-Net [51]	0.3125	0.2575	0.3889	0.2111
CO Solver	0.5750	0.5250	0.5333	0.4667
LLMCO4MR (Ours)	**0.6750**	0.6250	**0.6222**	0.5556

Ablation Study. We also present ablation study results in Tab. 5. "CO Solver" refers to deriving matching pairs directly from the output of CO solver. Few selected MLLMs exhibit capabilities in understanding cross-image fragment relations. These MLLMs' performances were evaluated by processing pairs of fragments without CO solver filtering out outliers. We conclude that similar to retrieval-augmented LLM frameworks, CO solver in our LLMCO4MR is a non-trivial extension, empowers MLLM capability handling combinatorial explosion problems. We detail in the Appendix the current limitations of MLLMs in directly addressing manuscript restoration tasks with combinatorial explosion.

Our multifaceted approach serves two purposes: it complements and enhances traditional contour-based matching, and it helps to stabilize the variable performance of MLLMs in image matching. As a result, our method demonstrates increased effectiveness in handling real-world manuscript restoration tasks, especially in scenarios where either matching strategy alone is insufficient.

5 Conclusion

We have presented a novel pipeline, LLMCO4MR, designed for the intricate task of restoring ancient manuscripts from fragments. We first formalize the problem as a cardinality-constrained combinatorial problem and then a neural combinatorial solver is devised to produce the Top-K candidates for merging. Then the MLLM is employed, leveraging the candidates as prompt for final restoration. Experiments on both synthetic fragments and real-world Dunhuang data demonstrate the potential of our approach, compared with the traditional approaches which require labors of well-trained archaeologists.

Limitation: Like existing works, our current approach also cannot handle the case with missing fragments and the fragments are required to be basically in a grid layout. Another unsolved case is the existence of both small and large fragments, which we leave for future work. Also, more case studies on broader scope need to be studied when such data becomes available.

Acknowledgements

This work was in part supported by National Science and Technology Major Project (2022ZD0119100), National Natural Science Foundation of China (No. 62441605, 72342023), National Social Science Fund of China (No. 21BYY142, 14AZS001), Key Research and Development Program of Zhejiang Province (No. 2024C03270), the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010).

References

1. Abitbol, R., Shimshoni, I., Ben-Dov, J.: Machine learning based assembly of fragments of ancient papyrus. Journal on Computing and Cultural Heritage (JOCCH) **14**(3), 1–21 (2021)
2. Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., de Freitas, N.: Restoring and attributing ancient texts using deep neural networks. Nature **603**(7900), 280–283 (2022)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report (2023)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
5. Bengio, Y., Lodi, A., Prouvost, A.: Machine learning for combinatorial optimization: a methodological tour d'horizon. European Journal of Operational Research **290**(2), 405–421 (2021)
6. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. Spie (1992)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
8. Buchbinder, N., Feldman, M., Naor, J., Schwartz, R.: Submodular maximization with cardinality constraints. In: Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. pp. 1433–1452. SIAM (2014)
9. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
10. Derech, N., Tal, A., Shimshoni, I.: Solving archaeological puzzles. Pattern Recognition **119**, 108065 (2021)
11. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model (2023)
12. Galambos, I.: Dunhuang Manuscript Culture: End of the First Millennium, vol. 22. Walter de Gruyter GmbH & Co KG (2020)

13. da Gama Leitao, H.C., Stolfi, J.: A multiscale method for the reassembly of two-dimensional fragmented objects. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(9), 1239–1251 (2002)
14. Ge, J., Luo, H., Qian, S., Gan, Y., Fu, J., Zhang, S.: Chain of thought prompt tuning in vision language models (2023)
15. Guo, P.F., Chen, Y.H., Tsai, Y.D., Lin, S.D.: Towards optimizing with large language models. arXiv preprint arXiv:2310.05204 (2023)
16. Guo, Q., Wang, R., Guo, J., Li, B., Song, K., Tan, X., Liu, G., Bian, J., Yang, Y.: Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. arXiv preprint arXiv:2309.08532 (2023)
17. Hossieni, S.S., Shabani, M.A., Irandoust, S., Furukawa, Y.: Puzzlefusion: Unleashing the power of diffusion models for spatial puzzle solving. Advances in Neural Information Processing Systems **36** (2024)
18. Karalias, N., Loukas, A.: Erdos goes neural: an unsupervised learning framework for combinatorial optimization on graphs. Advances in Neural Information Processing Systems **33**, 6659–6672 (2020)
19. Le, C., Li, X.: Jigsawnet: Shredded image reassembly using convolutional neural network and loop-based composition. IEEE Transactions on Image Processing **28**(8), 4000–4015 (2019)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
22. Lu, J., Sun, Y., Huang, Q.: Jigsaw: Learning to assemble multiple fractured objects. arXiv preprint arXiv:2305.17975 (2023)
23. Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S., Tu, Z.: Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration (2023)
24. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models (2023)
25. Markaki, S., Panagiotakis, C.: Jigsaw puzzle solving techniques and applications: a survey. The Visual Computer pp. 1–17 (2022)
26. OpenAI: Gpt-4 technical report (2023)
27. Panagiotakis, C., Markaki, S., Kokinou, E., Papadakis, H.: Coastline matching via a graph-based approach. Computational Geosciences **26**(6), 1439–1448 (2022)
28. Paumard, M.M., Picard, D., Tabia, H.: Jigsaw puzzle solving using local feature co-occurrences in deep neural networks. In: 2018 25th IEEE international conference on image processing (ICIP). pp. 1018–1022. IEEE (2018)
29. Pengcheng, G., Gang, G., Jiangqin, W., Baogang, W.: Chinese calligraphic style representation for recognition. International Journal on Document Analysis and Recognition (IJDAR) **20**, 59–68 (2017)
30. Pirrone, A., Aimar, M.B., Journet, N.: Papy-s-net: A siamese network to match papyrus fragments. In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing. pp. 78–83 (2019)
31. Pirrone, A., Beurton-Aimar, M., Journet, N.: Self-supervised deep metric learning for ancient papyrus fragments retrieval. International Journal on Document Analysis and Recognition (IJDAR) **24**(3), 219–234 (2021)
32. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001)
33. Savelka, J., Ashley, K.D., Gray, M.A., Westermann, H., Xu, H.: Explaining legal concepts with augmented large language models (gpt-4) (2023)

34. Savino, P., Tonazzini, A.: Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs. Journal of Cultural Heritage **19**, 511–521 (2016)
35. Shuai, R.W., Ruffolo, J.A., Gray, J.J.: Generative language modeling for antibody design. bioRxiv (2022). https://doi.org/10.1101/2021.12.13.472419, https://www.biorxiv.org/content/early/2022/12/20/2021.12.13.472419
36. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B.A., Webster, D., Corrado, G.S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., Natarajan, V.: Large language models encode clinical knowledge (2022)
37. Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., de Freitas, N.: Machine learning for ancient languages: A survey. Computational Linguistics pp. 1–44 (2023)
38. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
39. Tsamoura, E., Pitas, I.: Automatic color based reassembly of fragmented images and paintings. IEEE Transactions on Image Processing **19**(3), 680–690 (2009)
40. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. Advances in neural information processing systems **28** (2015)
41. Wang, R., Shen, L., Chen, Y., Yang, X., Tao, D., Yan, J.: Towards one-shot neural combinatorial solvers: Theoretical and empirical notes on the cardinality-constrained case. In: The Eleventh International Conference on Learning Representations (2022)
42. Wang, R., Yan, J., Yang, X.: Learning combinatorial embedding networks for deep graph matching. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3056–3065 (2019)
43. Wang, R., Zhang, Y., Guo, Z., Chen, T., Yang, X., Yan, J.: Linsatnet: The positive linear satisfiability neural networks. In: International Conference on Machine Learning (ICML) (2023)
44. Wolfson, H., Schonberg, E., Kalvin, A., Lamdan, Y.: Solving jigsaw puzzles by computer. Annals of Operations Research **12**(1), 51–64 (1988)
45. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023)
46. Xie, Y., Dai, H., Chen, M., Dai, B., Zhao, T., Zha, H., Wei, W., Pfister, T.: Differentiable top-k with optimal transport. Advances in Neural Information Processing Systems **33**, 20520–20531 (2020)
47. Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., Yang, F., Deng, F., Wang, F., Liu, F., Ai, G., Dong, G., Zhao, H., Xu, H., Sun, H., Zhang, H., Liu, H., Ji, J., Xie, J., Dai, J., Fang, K., Su, L., Song, L., Liu, L., Ru, L., Ma, L., Wang, M., Liu, M., Lin, M., Nie, N., Guo, P., Sun, R., Zhang, T., Li, T., Li, T., Cheng, W., Chen, W., Zeng, X., Wang, X., Chen, X., Men, X., Yu, X., Pan, X., Shen, Y., Wang, Y., Li, Y., Jiang, Y., Gao, Y., Zhang, Y., Zhou, Z., Wu, Z.: Baichuan 2: Open large-scale language models (2023)
48. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v(ision) (2023)

49. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action (2023)
50. Yu, F., Quartey, L., Schilder, F.: Legal prompting: Teaching a language model to think like a lawyer (2022)
51. Zhang, C., Wang, B., Chen, K., Zong, R., Mo, B.f., Men, Y., Almpanidis, G., Chen, S., Zhang, X.: Data-driven oracle bone rejoining: A dataset and practical self-supervised learning scheme. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4482–4492 (2022)
52. Zhang, C., Zong, R., Cao, S., Men, Y., Mo, B.: Ai-powered oracle bone inscriptions recognition and fragments rejoining. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. pp. 5309–5311 (2021)
53. Zhang, K., Li, X.: A graph-based optimization algorithm for fragmented image reassembly. Graphical Models **76**(5), 484–495 (2014)
54. Zhang, Y., Fang, Z., Yang, X., Zhang, S., He, B., Dou, H., Yan, J., Zhang, Y., Wu, F.: Reconnecting the broken civilization: Patchwork integration of fragments from ancient manuscripts. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 1157–1166 (2023)
55. Zhang, Z., Wang, Y.T., Li, B., Guo, A., Liu, C.L.: Deep rejoining model for oracle bone fragment image. In: Asian Conference on Pattern Recognition. pp. 3–15. Springer (2021)
56. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023)