SHSP: STRUCTURE-AWARE HIERARCHICAL SOLUTION PREDICTION FOR MIXED-INTEGER LINEAR PROGRAMMING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed-Integer Linear Programming (MILP) is a fundamental optimization paradigm in combinatorial optimization and has been widely applied across real-world domains. Due to its NP-hard nature, obtaining optimal solutions for large-scale or highly constrained MILP instances remains computationally prohibitive. Learning-based solution prediction has therefore emerged as a promising approach to provide high-quality variable assignments for solver acceleration. However, existing methods typically adopt a one-shot prediction paradigm that predicts the marginal probabilities of all variables simultaneously. As a result, the conditional dependencies among variables are only implicitly captured through message passing, with the burden of modeling the combinatorial structure falling entirely on the representational capacity of graph neural networks. To address this limitation, we propose the **S**tructure-Aware **H**ierarchical **S**olution **P**rediction (SHSP) framework that replaces the parallel marginal decoding of one-shot methods with a novel hierarchical conditional decoding mechanism. Specifically, SHSP constructs a variable coupling graph from the constraint structure, decodes variables sequentially along a hierarchy of increasing coupling strength, and conditions each hierarchy on previously predicted assignments. To mitigate error accumulation during the decoding process, SHSP further incorporates a confidence-aware mask-and-repair mechanism to identify and correct unreliable intermediate predictions. We integrate SHSP with multiple learning-guided search methods, and evaluate it on four standard MILP benchmarks. Experimental results demonstrate that SHSP significantly outperforms existing one-shot prediction baselines, achieving a 54% average reduction in solution gap.

1 INTRODUCTION

Mixed-Integer Linear Programming (MILP) is one of the most widely used modeling paradigms in combinatorial optimization and operations research, with broad applications in various optimization domains (Ma et al., 2019; Li et al., 2024b). Owing to the NP-hardness of MILP (Karp, 1972), decades of research efforts have been devoted to developing powerful solvers such as SCIP (Achterberg, 2009) and Gurobi (Gurobi Optimization, LLC, 2026). Equipped with advanced techniques including cutting planes, primal heuristics, and presolving, these solvers are capable of efficiently handling a wide range of practical instances. However, solving large-scale and highly constrained instances remains computationally expensive, making the improvement of MILP solving efficiency a fundamental research challenge.

In recent years, machine learning has emerged as a promising paradigm for accelerating MILP solving (Li et al., 2024a), and existing studies generally fall into two categories. The first category learns heuristic policies for key components of branch-and-bound (B&B) solvers, such as branching (Khalil et al., 2016; Gasse et al., 2019; Hou et al., 2026) and cut selection strategies (Tang et al., 2020; Huang et al., 2021; Paulus et al., 2022; M. et al., 2026), typically via imitation learning or reinforcement learning. The second category employs machine learning models such as graph neural networks (GNNs) to directly predict high-quality solutions for MILP instances (Ding et al., 2020; Zeng et al., 2024; Geng et al., 2025). Research in this category advances along two complementary aspects. The first aspect concerns how predictions are exploited to reduce the search space of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

original problem (Nair et al., 2020; Han et al., 2023; Liu et al., 2025). The second aspect focuses on improving the prediction capacity (Pu et al., 2026; Wang et al., 2026). A detailed discussion of related work is provided in Appendix A. Together, these efforts have established learning-based solution prediction as an effective paradigm for accelerating MILP solving.

Despite the remarkable performance of existing solution prediction methods, most of them follow a one-shot prediction paradigm (Han et al., 2023; Liu et al., 2025; Pu et al., 2026), where the marginal probabilities of all decision variables are predicted simultaneously in a single forward pass. However, variables in a MILP are strongly interdependent through shared constraints. The optimal assignment of one variable often hinges on the combinatorial structure formed by the others, to the extent that exactly predicting the complete optimal assignments would amount to solving the MILP itself. Under the one-shot paradigm, such dependencies are captured only implicitly by the GNN encoder, while the decoding stage treats all variables as independent. This mismatch can yield mutually conflicting assignments that mislead, rather than accelerate, the downstream solver.

To address these limitations, we propose **Structure-Aware Hierarchical Solution Prediction for MILP (SHSP)**, a novel learning framework that explicitly models variable dependencies through hierarchical partitioning and prediction, rather than relying on a one-shot predictor to recover them implicitly. Specifically, SHSP first constructs a weighted variable coupling graph that quantifies pairwise dependencies based on expected constraint violation and coefficient importance. Guided by the coupling scores, variables are predicted in a hierarchical manner: weakly coupled variables are resolved first, while strongly coupled ones are subsequently decoded conditioned on the resulting partial assignments. To further enhance robustness, we introduce a confidence-based mask-and-repair mechanism that mitigates error accumulation during multi-step inference. Finally, the predicted assignments are exploited through a structure-aware fixing strategy that prioritizes strongly coupled variables, whose fixed values propagate through shared constraints and tighten the feasible domains of many others, thereby effectively reducing the search space.

As a superior predictor, SHSP serves as a drop-in replacement for the vanilla GNN-based one-shot predictor and can be seamlessly integrated into diverse learning-guided acceleration frameworks (Nair et al., 2020; Han et al., 2023). To validate the effectiveness of SHSP, we conduct extensive experiments on four widely used MILP benchmarks. Experimental results demonstrate that SHSP consistently outperforms the one-shot prediction baselines across all benchmarks, reducing the average absolute primal gap by up to 100.0%, 38.1%, 94.4%, and 42.1%, respectively. Notably, a variant of SHSP even surpasses the solution quality of full-budget Gurobi on combinatorial auction instances within a substantially smaller time budget.

2 PRELIMINARIES

2.1 MIXED-INTEGER LINEAR PROGRAMMING

Mixed-integer linear programming (MILP) seeks to optimize a linear objective function over a feasible region defined by linear constraints, where a subset of decision variables take integer values. Formally, an MILP instance can be expressed as

$$\begin{aligned} \min_{x} \quad & c^\top x \\ \text{s.t.} \quad & Ax \le b, \\ & l \le x \le u, \\ & x \in \mathbb{Z}^p \times \mathbb{R}^{n-p}, \end{aligned}$$

where x denotes the n-dimensional decision vector comprising p integer variables and $n-p$ continuous variables. The vector $c \in \mathbb{R}^n$ specifies the objective coefficients, $A \in \mathbb{R}^{m\times n}$ is the constraint coefficient matrix, and $b \in \mathbb{R}^m$ is the corresponding right-hand-side vector. The vectors $l \in (\mathbb{R} \cup \{-\infty\})^n$ and $u \in (\mathbb{R} \cup \{+\infty\})^n$ impose the lower and upper bounds on the variables, respectively. Following prior works (Han et al., 2023; Liu et al., 2025; Pu et al., 2026), we focus on MILP instances whose integer variables are binary, as general integer variables can be reduced to binary ones via well-established preprocessing techniques (Nair et al., 2020).

An MILP instance can be naturally encoded as a weighted bipartite graph $G = (W \cup V, E)$ (Gasse et al., 2019), where W and V denote the sets of constraint nodes and variable nodes, respectively.

An edge $(w, v) \in E$ connects a constraint node $w \in W$ and a variable node $v \in V$ if and only if the corresponding variable appears with a nonzero coefficient in the constraint. The node features typically encompass objective coefficients, variable bounds, and variable types on the variable side, as well as right-hand-side values and constraint senses on the constraint side, while the edge features carry the corresponding constraint coefficients. More details on the graph modeling procedure are provided in Appendix B.

2.2 PREDICT-AND-SEARCH

Predict-and-Search (PaS) (Han et al., 2023) is a two-stage framework that integrates neural solution prediction with mathematical optimization for accelerating MILP solving. Given an MILP instance I, PaS first learns the underlying solution distribution from a collection of optimal or near-optimal solutions. Specifically, the solution distribution is approximated as $q(x|I) = \frac{\exp(-E(x,I))}{\sum_{x' \in S} \exp(-E(x',I))}$, where S denotes the collected solution set and the energy function is defined as $E(x, I) = c^\top x$ if x is feasible and $+\infty$ otherwise. Guided by this distribution, PaS trains a graph neural network p_θ to estimate the marginal probability of each binary variable. Under the variable-wise independence assumption, the joint solution distribution is factorized as $p_\theta(x \mid I) = \prod_{i=1}^{n} p_\theta(x_i \mid I)$, thereby reducing the learning task to predicting per-variable marginals. The GNN is trained by minimizing a weighted binary cross-entropy loss over the collected solutions, where each solution is weighted according to $q(x \mid I)$:

$$\mathcal{L} = \sum_{s=1}^{|S|} q(x^{(s)} \mid I) \sum_{i=1}^{n} \ell(\hat{p}_i, y_{s,i})$$

where $\hat{p}_i = p_\theta(x_i = 1 \mid I)$ denotes the predicted probability of variable x_i, $y_{s,i} \in \{0, 1\}$ denotes the value of variable x_i in solution $x^{(s)}$, and $\ell(\hat{p}_i, y_{s,i}) = -y_{s,i} \log \hat{p}_i - (1 - y_{s,i}) \log(1 - \hat{p}_i)$ denotes the binary cross-entropy loss. Based on the predicted marginal probabilities, PaS selects k_1 variables with the highest marginal probabilities and fixes them to one, while selecting k_0 variables with the lowest marginal probabilities and fixing them to zero. The resulting partial solution is denoted as $\hat{x}^{[P]}$, where P represents the index set of fixed variables.

Rather than hard-fixing these variables as in neural diving (Nair et al., 2020), PaS formulates a trust-region optimization problem that constrains the feasible region to the neighborhood of the predicted partial solution. Formally, the trust region is defined as $B_P(\hat{x}^{[P]}, \Delta) = \{x^{[P]} \in \mathbb{R}^{|P|} \mid \|\hat{x}^{[P]} - x^{[P]}\|_1 \leq \Delta\}$, where Δ denotes the neighborhood radius. The resulting trust-region problem is formulated as

$$\begin{aligned} \min_{x} \quad & c^\top x \\ \text{s.t.} \quad & Ax \leq b, \quad l \leq x \leq u, \\ & x^{[P]} \in B_P(\hat{x}^{[P]}, \Delta), \\ & x \in \mathbb{Z}^p \times \mathbb{R}^{n-p}. \end{aligned}$$

By coupling neural prediction with trust-region-based local optimization, PaS effectively combines the efficiency of learning-based methods and the reliability of mathematical programming solvers. Notably, when $\Delta = 0$, PaS degenerates to the hard-fixing scheme of neural diving.

3 METHODOLOGY

In this section, we present **Structure-Aware Hierarchical Solution Prediction for MILP (SHSP)**. In contrast to existing methods that predict the assignments of all decision variables simultaneously, SHSP explicitly exploits variable coupling information and decodes variable assignments sequentially following a hierarchy of increasing coupling strength, where the prediction at each decoding step is conditioned on the partial assignments of lower hierarchies obtained in preceding steps. The overall framework is illustrated in Figure 1. In the remainder of this section, we first introduce the variable coupling graph, then describe the structure-aware hierarchical predictor, and finally present the corresponding variable fixing strategy.

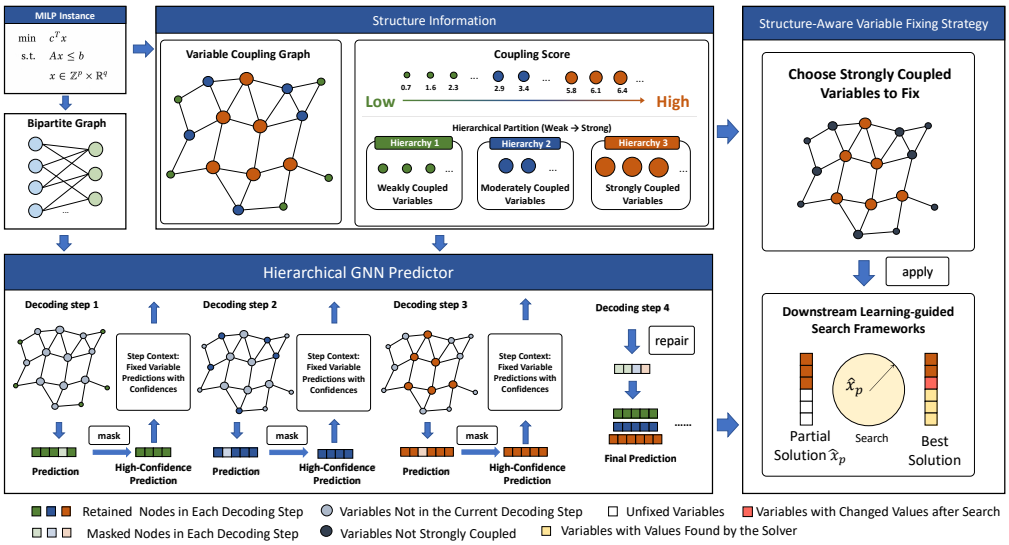

Figure 1: Overview of the proposed method. Given an original MILP instance, we first construct a variable coupling graph to capture the structural dependencies among decision variables. Based on the variable coupling graph, variables are ranked according to their coupling score and divided into multiple hierarchies. Our SHSP sequentially predicts variables following the weak-to-strong hierarchy, with a mask-and-repair mechanism to mitigate error accumulation. Finally, the predicted variable assignments are combined with a structure-aware variable fixing strategy and applied to downstream learning-guided search frameworks such as Predict-and-Search (Han et al., 2023).

3.1 Variable Coupling Graph

Decision variables are often tightly coupled through shared constraints, and such structural dependencies play a critical role in obtaining high-quality solutions. To explicitly capture these dependencies, we construct a variable coupling graph, which serves as the basis for estimating coupling scores and for determining hierarchical partitioning.

Graph Modeling. We model the coupling structure among variables as an undirected weighted graph, where nodes correspond to variables and edges capture their co-occurrence in constraints. Given linear constraints, we consider all variables when estimating coupling strengths, but retain only those that co-occur with at least one binary variable in some constraints, i.e., those directly connected to binary variables. This makes the coupling graph considerably sparser without severely compromising the prediction accuracy of binary variables. An edge is created between two retained variables if they appear together in at least one constraint, with its weight quantifying the coupling strength implied by the shared constraints, as detailed below.

Edge Weight Computation. For each retained variable pair, we assign an edge weight that reflects the coupling strength implied by the constraints. The weight is determined by two complementary heuristic factors. The first factor, termed the *expected violation*, measures how strongly a variable pair is restricted by a constraint. Specifically, we treat the two variables as random variables, where binary variables follow independent uniform distributions over $\{0, 1\}$ and continuous variables follow uniform distributions over their bounds. We then compute the expected violation of the constraint induced by the pair alone. A larger expected violation indicates that the pair is more tightly restricted by the constraint and thus more strongly coupled. The second factor, termed the *coefficient importance*, characterizes the relative contribution of a variable pair within a constraint. It is computed as the sum of the magnitudes of the two corresponding coefficients, each scaled by the range of its associated variable. To eliminate scale discrepancies across constraints, both factors are normalized by their respective average values over all variable pairs within the same constraint. It is noteworthy that both factors are computed over all variables, including purely continuous ones.

The local coupling weight of a pair in constraint c is then given by the product of the two normalized factors, and the final edge weight W_{ij} between variables z_i and z_j is obtained by summing the local weights over all constraints containing both variables. The detailed computation of edge weights is provided in Appendix C. The resulting weighted graph captures the overall structural coupling among variables and serves as the basis for computing the variable coupling score and determining the hierarchical prediction order in structure-aware hierarchical prediction.

3.2 Structure-Aware Hierarchical Prediction

The above weighted variable coupling graph provides a quantitative measure of the structural dependency among variables. Building upon this, we design a structure-aware hierarchical prediction framework that decodes variables sequentially from weakly coupled to strongly coupled ones. This framework conditions each decoding step on the assignments fixed in preceding hierarchies, so that strongly coupled variables, whose values are the most sensitive to those of others, are predicted with explicit structural context rather than in isolation as in previous one-shot methods.

Coupling Score-Based Hierarchical Partitioning. For each binary variable z_i, we define its variable coupling score (VCS) as the sum of the weights of its incident edges in the coupling graph, which quantifies the overall strength of its structural dependency on the remaining variables. All variables are sorted in ascending order of VCS and evenly partitioned into K hierarchies $\{H_1, H_2, \ldots, H_K\}$, such that H_1 contains the most weakly coupled variables and H_K the most strongly coupled ones. The decoding process then follows this weak-to-strong order. Weakly coupled variables are largely insensitive to the assignments of other variables and can thus be predicted reliably at early stages. Their fixed values in turn serve as conditioning information that anchors the prediction of strongly coupled variables in later hierarchies.

Hierarchical Decoding with Mask-and-Repair. The model decodes the K hierarchies sequentially, and each step is conditioned on the tentative assignments accumulated from all preceding hierarchies. To realize this conditioning, the input feature of every variable node is augmented with a decoding-state triplet $(m_i, \hat{x}_i, c_i)$, where $m_i \in \{0, 1\}$ indicates whether variable i should be predicted in the current step, $\hat{x}_i$ denotes its tentative binary assignment, and c_i its associated confidence. For variables that will be decoded in future steps, the triplet is set to a zero placeholder. In addition, to make the network aware of its position in the decoding process, the current step index is encoded via positional encoding (Vaswani et al., 2017), and the resulting step embedding is concatenated with the initial node embeddings and passed through a fusion layer to produce the step-aware node representations. The collection of these state features constitutes the conditioning context $\mathbf{h}_{<k}$, and the prediction of the k-th hierarchy is formulated as

$$\mathbf{p}_{H_k} = f_\theta(G, H_k, \mathbf{h}_{<k}),$$

where G denotes the bipartite graph and f_θ the graph neural network. For each variable $i \in H_k$, the network outputs a probability p_i, from which the binary assignment $\hat{x}_i = \mathbb{I}(p_i > 0.5)$ and the confidence $c_i = 2|p_i - 0.5|$ are derived.

Since the conditioning context is constructed from model predictions rather than ground-truth assignments, wrong early decisions may propagate and accumulate across hierarchies. To mitigate such error accumulation, we propose a mask-and-repair mechanism. Before being incorporated into the conditioning context, each newly predicted variable is stochastically masked with a probability $p_{\text{mask}}(i) \propto 1 - c_i$ that decreases with its confidence. Consequently, high-confidence predictions are likely to be retained as conditioning information, while uncertain predictions tend to be masked, reverting to the undecided placeholder state and being collected into a repair set $\mathcal{R}$. After all K hierarchies have been decoded, a final repair step decodes the variables in $\mathcal{R}$ conditioned on the full context of retained assignments, i.e., $\mathbf{p}_{\mathcal{R}} = f_\theta\big(G, \mathcal{R}, \mathbf{h}_{\mathcal{V}\setminus\mathcal{R}}\big)$, and the repaired predictions replace the masked ones to yield the final solution. This iterative masking-and-prediction paradigm shares a similar spirit with the recent work Apollo-MILP (Liu et al., 2025). However, our method differs in two key aspects: (1) it explicitly exploits the variable coupling structure to establish a hierarchy that guides the decoding process; (2) it constitutes an inherently new prediction paradigm that operates entirely within the prediction stage, rather than following Apollo-MILP's alternation between prediction and solver invocation.

Training Objective. The framework is trained by minimizing a weighted prediction loss aggregated over all hierarchies and the repair step. Given S supervised solutions with objective values $\{o_s\}_{s=1}^{S}$, each solution is assigned a weight $w_s = \exp(-o_s/\tau) \,/\, \sum_{j=1}^{S} \exp(-o_j/\tau)$, where higher-quality solutions contribute more to the supervision and τ controls the weighting smoothness. The overall training objective is defined as

$$\mathcal{L} = \frac{1}{K+1}\left(\sum_{k=1}^{K}\sum_{s=1}^{S} w_s \sum_{i \in H_k} \ell\Big(p_i^{(k)}, y_{s,i}\Big) + \sum_{s=1}^{S} w_s \sum_{i \in \mathcal{R}} \ell\Big(p_i^{(\mathrm{rep})}, y_{s,i}\Big) \right),$$

where $p_i^{(k)}$ and $p_i^{(\mathrm{rep})}$ denote the predicted probabilities of variable i in the k-th hierarchy and the repair step, respectively, $y_{s,i}$ denotes the label of variable i in solution s, and $\ell(\cdot,\cdot)$ is the binary cross-entropy loss. Note that when computing the k-th prediction, the preceding $k-1$ states are decoded by the model itself, which aligns with the inference procedure and thus mitigates the distributional shift between training and inference. However, fully relying on model-decoded states can hinder convergence, as early predictions are highly inaccurate. We therefore incorporate a *teacher forcing* strategy that employs ground-truth assignments as conditioning information in the early stage of training, and the proportion of teacher forcing is progressively decayed as training proceeds (see details in Appendix E.2).

3.3 Structure-Aware Variable Fixing Strategy

Existing variable fixing strategies typically select variables solely according to their prediction confidences. However, fixing variables with similar confidence levels may have substantially different structural influences on the MILP problem. A strongly coupled variable is connected to more variables through shared constraints than a weakly coupled variable. Fixing its value can therefore restrict the feasible values of those connected variables, leading to a greater reduction in the feasible region. Therefore, variable fixing should consider not only prediction confidence but also the structural influence of each variable.

Motivated by this observation, we propose a structure-aware variable fixing strategy that jointly considers the variable coupling structure and the prediction confidences. Specifically, we first identify variables whose variable coupling scores exceed a prescribed threshold as high-coupling variables. We then fix the selected variables to their predicted values only if their prediction confidence is sufficiently high. This two-stage screening ensures that only strongly coupled variables with highly reliable predictions are eligible for fixing, thereby enabling downstream optimization frameworks to explicitly exploit structural information for more effective search-space reduction. Details of the structure-aware variable fixing strategy are provided in Appendix E.3.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate our framework on four widely adopted MILP benchmarks in this field: Combinatorial Auctions (CA) (Leyton-Brown et al., 2000), Workload Apportionment (WA) (Gasse et al., 2022), Item Placement (IP) (Gasse et al., 2022), and Set Covering (SC) (Balas & Ho, 1980). Following existing studies (Gasse et al., 2019; Han et al., 2023; Liu et al., 2025), we generate SC and CA instances using similar procedures, while the IP and WA benchmarks are obtained from two real-world problems used in the NeurIPS ML4CO 2021 competition (Gasse et al., 2022). The statistical information is provided in Table 1. For each benchmark problem, we use 240 instances for training, 60 instances for validation, and 100 instances for testing.

Table 1: Summary statistics of the benchmark datasets used in the main evaluation.

	CA	WA	IP	SC
Constraint Number	2592	64318	195	3000
Variable Number	1500	61000	1083	5000
Binary Variables	1500	1000	1050	5000
Continuous Variables	0	60000	33	0
Integer Variables	0	0	0	0

Baselines We compare our method with both traditional MILP solvers and learning-based solution prediction approaches. For traditional optimization methods, we compare our method with

Table 2: Performance comparison on four benchmark datasets (CA, WA, IP, and SC) using Gurobi under a 1000-second time limit. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. Bold values denote the better results. Improvement reports the absolute improvement in Obj and the relative reduction in Gap_{abs} compared with the corresponding original solution-prediction framework.

Method		CA ↑ (BKS:98489.24)		WA ↓ (BKS:706.86)		IP ↓ (BKS:11.69)		SC ↓ (BKS:123.30)	
		Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}
Gurobi (3600s)		98453.42	35.82	706.86	0.00	11.69	0.00	123.30	0.00
Gurobi (1000s)		97407.51	1081.73	707.07	0.21	12.16	0.47	123.54	0.24
ND	Original	94340.63	4148.61	707.12	0.26	14.18	2.49	123.51	0.21
	SHSP	**97121.05**	**1368.19**	**707.07**	**0.21**	**11.83**	**0.14**	123.51	0.21
Improvement		2780.42	67.02%	0.05	19.23%	2.35	94.38%	0.00	0.00%
PaS	Original	97891.74	597.50	707.07	0.21	12.06	0.37	123.49	0.19
	SHSP	**98487.72**	1.52	**706.99**	**0.13**	**11.97**	0.28	**123.41**	0.11
Improvement		595.98	99.75%	0.08	38.1%	0.09	24.32%	0.08	42.11%
Apollo	Original	97997.74	491.50	707.07	0.21	11.97	0.28	123.42	0.12
	SHSP	**98489.24**	**0.00**	**707.01**	**0.15**	**11.71**	**0.02**	**123.37**	**0.07**
Improvement		491.50	100.00%	0.06	28.6%	0.26	92.86%	0.05	41.67%

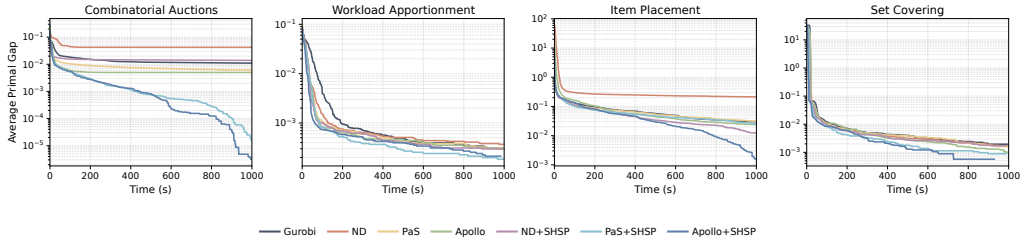

Figure 2: The convergence curves of primal gaps as the solving process proceeds. All methods are implemented using Gurobi with a total time budget of 1,000 seconds. For a fair comparison, the graph construction time of SHSP is included in this budget. Results are averaged over 100 testing instances. A lower primal gap indicates faster convergence and better solving performance.

Gurobi (Gurobi Optimization, LLC, 2026) and SCIP (Achterberg, 2009). For learning-based approaches, we consider three representative methods: Neural Diving (ND) (Nair et al., 2020), Predict-and-Search (PaS) (Han et al., 2023), and Apollo-MILP (Liu et al., 2025).

Metrics. We compare the best objective values (OBJ) achieved by different methods within a 1000-second time limit on each test instance. Following the evaluation protocol in Predict-and-Search (Han et al., 2023), we use the best solution obtained by a single-threaded Gurobi solver with a 3600-second time limit as the best-known solution (BKS), which serves as an approximation of the optimal value. The absolute primal gap is then calculated as $\text{Gap}_{abs} = |\text{OBJ} - \text{BKS}|$, measuring the distance between the obtained solution and the BKS. It is worth noting that our method achieves better solutions than the Gurobi solver with a 3600-second time limit on certain benchmark problems. In such cases, the objective values found by our method are used to update the BKS.

4.2 Main Evaluation

Solving Performance To evaluate the effectiveness of our method, we replace the original GNN predictor and the variable fixing strategy in each framework with our SHSP predictor and structure-aware variable fixing strategy, forming **ND+SHSP**, **PaS+SHSP**, and **Apollo+SHSP**, respectively.

Table 3: Performance comparison on four benchmark datasets (CA, WA, IP, and SC) using SCIP under a 1000-second time limit. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. Bold values denote the better results. Improvement reports the absolute improvement in Obj and the relative reduction in Gap_{abs} compared with the corresponding downstream solution-prediction framework.

Method		CA ↑ (BKS:98489.24)		WA ↓ (BKS:706.86)		IP ↓ (BKS:11.69)		SC ↓ (BKS:123.30)	
		Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}
SCIP (3600s)		97795.78	693.46	708.21	1.35	19.06	7.37	124.58	1.28
SCIP (1000s)		96932.37	1556.87	709.62	2.76	22.69	11.00	125.81	2.51
ND	Original	94338.51	4150.73	707.83	0.97	16.33	4.64	125.87	2.57
	SHSP	**97267.46**	**1221.78**	**707.18**	**0.32**	**13.70**	**2.01**	**125.84**	**2.54**
Improvement		2928.95	70.57%	0.65	67.01%	2.63	56.68%	0.03	1.17%
PaS	Original	97223.78	1265.46	708.27	1.41	19.61	7.92	125.82	2.52
	SHSP	**97230.19**	**1259.05**	**708.21**	**1.35**	**19.44**	**7.75**	**125.33**	**2.03**
Improvement		6.41	0.51%	0.06	4.26%	0.17	2.15%	0.49	19.44%
Apollo	Original	96270.42	2218.82	708.42	1.56	20.48	8.79	125.74	2.44
	SHSP	**97192.31**	**1296.93**	**708.21**	**1.35**	**18.56**	**6.87**	**125.45**	**2.15**
Improvement		921.89	41.55%	0.21	13.46%	1.92	21.84%	0.29	11.89%

We compare the solving performance between our framework and the baselines under a time limit of 1,000 seconds. Although SHSP performs hierarchical multi-step prediction, its additional neural network inference time is less than 0.2 seconds across all benchmark datasets. The graph construction time ranges from 0.26 to 6.77 seconds and is included in the 1,000-second time budget for a fair comparison. Table 2 presents the average objective values and the corresponding average absolute primal gaps across four benchmark datasets, with Gurobi employed as the downstream solver. Similar results obtained with the SCIP solver are reported in Table 3. Detailed results of graph construction time and neural network inference time are provided in Appendix F.3.

Overall, SHSP consistently improves performance across all three learning-guided search frameworks compared with their original one-shot prediction baselines. With PaS, SHSP reduces the absolute primal gap by 99.7% on CA, nearly closing the gap entirely, along with reductions of 38.1%, 24.3%, and 42.1% on WA, IP, and SC, respectively. With Apollo, the gains are similarly substantial: the primal gap is completely eliminated on CA (100.0% reduction) and reduced by 92.9% on IP, with further reductions of 28.6% and 41.7% on WA and SC. When integrated with ND, SHSP also delivers solid improvements, reducing the absolute primal gap by 94.4% on IP and 67.0% on CA, and by 19.2% on WA. Remarkable improvements are also observed when using SCIP as the downstream solver. As shown in Table 3, SHSP consistently reduces the absolute primal gap across all three frameworks on all four benchmarks, demonstrating that the proposed method generalizes well across different downstream solvers. It is worth noting that Apollo+SHSP reaches the best known solution, surpassing the solution obtained by Gurobi with a 3,600-second time limit on the CA benchmark. This observation suggests that our method can provide a stronger initialization for the downstream solver, not only accelerating the search process but also guiding it toward higher-quality solutions.

Primal Gap as a Function of Runtime Figure 2 presents the curves of the average primal gap, defined as $\text{Gap}_{\text{rel}} = \frac{|\text{OBJ}-\text{BKS}|}{|\text{BKS}|}$, throughout the solving process. Overall, SHSP-based methods consistently exhibit faster convergence and achieve lower primal gaps on the benchmarks than corresponding one-shot prediction methods. This behavior is attributed to more accurate solution predictions and more effective search-space reduction of our method.

Table 4: Ablation study of the proposed confidence-based mask-and-repair strategy on the CA and IP datasets. Bold values denote the better results.

Method	CA ↑	IP ↓
PaS+SHSP (None)	98380.92	12.12
PaS+SHSP (Mask)	98416.55	12.17
PaS+SHSP (Mask+Repair)	**98487.72**	**11.97**
Apollo+SHSP (None)	98385.20	11.78
Apollo+SHSP (Mask)	98344.07	11.80
Apollo+SHSP (Mask+Repair)	**98489.24**	**11.71**

Table 5: Ablation study of the proposed hierarchical solution prediction and structure-aware fixing strategy on the CA and IP datasets. Bold values denote the better results.

Method (Fixing Strategy)	CA ↑	IP ↓
PaS (Structure-Aware)	97184.33	12.12
PaS+SHSP (Original)	98061.51	12.34
PaS+SHSP (Structure-Aware)	**98487.72**	**11.97**
Apollo (Structure-Aware)	98015.32	11.87
Apollo+SHSP (Original)	98070.43	12.08
Apollo+SHSP (Structure-Aware)	**98489.24**	**11.71**

4.3 Ablation Study

Confidence-Based Mask and Repair Mechanism To evaluate the effectiveness of the proposed mask-and-repair mechanism, we compare three settings: without mask and repair (None), with the mask stage only (Mask), and with both mask and repair (Mask+Repair). As shown in Table 4, applying the masking mechanism alone does not consistently improve performance and may even cause slight degradation, since masked variables are simply discarded without correction, resulting in a loss of predictive information. In contrast, coupling masking with the proposed repair strategy consistently yields the best results across all datasets and both prediction frameworks, confirming that the two components are complementary and jointly contribute to the overall performance. More results are provided in Appendix F.1.

Hierarchical Solution Prediction Predictor To evaluate the effectiveness of the proposed hierarchical solution prediction predictor, we compare it with the original GNN predictor under the same structure-aware fixing strategy. As shown in Table 5, replacing the original GNN predictor with the hierarchical solution prediction predictor consistently improves the solution quality on both the PaS and Apollo frameworks. The improvement suggests that our predictor generates more accurate solution predictions, thereby enabling more effective variable fixing and subsequent problem reduction. Results on other problems are provided in Appendix F.1.

Structure-Aware Fixing Strategies To evaluate the effectiveness of the proposed structure-aware fixing strategy, we compare it with the original fixing strategy while using the same predictor. As shown in Table 5, the proposed structure-aware fixing strategy consistently outperforms the normal fixing strategy on both the PaS and Apollo frameworks. The improvement suggests that prioritizing structurally important variables leads to more effective search space reduction, allowing the solver to explore the remaining search space more efficiently and ultimately achieve better solving performance. Results on other problems are provided in Appendix F.1.

5 Conclusion

In this paper, we propose SHSP, a structure-aware hierarchical solution prediction framework for MILP solving. SHSP replaces the parallel marginal decoding with a hierarchical conditional decoding mechanism guided by a variable coupling graph, decoding variables from weakly to strongly coupled ones while conditioning each step on previously predicted assignments. It further incorporates a mask-and-repair mechanism to mitigate error accumulation, and a structure-aware fixing strategy to enable more effective search-space reduction in downstream frameworks. Extensive experiments on four standard benchmarks show that SHSP consistently outperforms one-shot baselines across three frameworks and two downstream solvers, achieving a 54% average reduction in the absolute primal gap with negligible inference overhead. In future work, we plan to learn coupling scores in an end-to-end manner and extend to broader problem classes.

REFERENCES

Tobias Achterberg. Scip: Solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41, 2009.

Egon Balas and Andrew Ho. *Set Covering Algorithms Using Cutting Planes, Heuristics, and Subgradient Optimization: A Computational Study*. Springer, 1980.

Antonia Chmiela, Elias Khalil, Ambros Gleixner, Andrea Lodi, and Sebastian Pokutta. Learning to schedule heuristics in branch and bound. In *Advances in Neural Information Processing Systems 34*, pp. 24235–24246, Virtual, 2021.

Jian-Ya Ding, Chao Zhang, Lei Shen, Shengyin Li, Bing Wang, Yinghui Xu, and Le Song. Accelerating primal solution findings for mixed integer programs based on solution prediction. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pp. 1452–1459, New York, NY, 2020.

Maxime Gasse, Didier Chetelat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact combinatorial optimization with graph convolutional neural networks. In *Advances in Neural Information Processing Systems 32*, pp. 15554–15566, Vancouver, Canada, 2019.

Maxime Gasse, Simon Bowly, Quentin Cappart, Jonas Charfreitag, Laurent Charlin, Didier Chetelat, Antonia Chmiela, Justin Dumouchelle, Ambros Gleixner, Aleksandr M. Kazachkov, Elias Khalil, Pawel Lichocki, Andrea Lodi, Miles Lubin, Chris J. Maddison, Morris Christopher, Dimitri J. Papageorgiou, Augustin Parjadis, Sebastian Pokutta, Antoine Prouvost, Lara Scavuzzo, Giulia Zarpellon, Linxin Yang, Sha Lai, Akang Wang, Xiaodong Luo, Xiang Zhou, Haohan Huang, Shengcheng Shao, Yuanming Zhu, Dong Zhang, Tao Quan, Zixuan Cao, Yang Xu, Zhewei Huang, Shuchang Zhou, Binbin Chen, Minggui He, Hao Hao, Zhiyu Zhang, Zhiwu An, and Kun Mao. The machine learning for combinatorial optimization competition (ml4co): Results and insights. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, pp. 220–231, Virtual, 2022.

Zijie Geng, Jie Wang, Xijun Li, Fangzhou Zhu, Jianye Hao, Bin Li, and Feng Wu. Differentiable integer linear programming. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, 2025.

Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp Christophel, Kati Jarck, Thorsten Koch, Jeff Linderoth, et al. Miplib 2017: Data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3):443–490, 2021.

Prateek Gupta, Maxime Gasse, Elias Khalil, Pawan Mudigonda, Andrea Lodi, and Yoshua Bengio. Hybrid models for learning to branch. In *Advances in Neural Information Processing Systems 33*, pp. 18087–18097, Virtual, 2020.

Prateek Gupta, Elias B. Khalil, Didier Chetelat, Maxime Gasse, Yoshua Bengio, Andrea Lodi, and M. Pawan Kumar. Lookback for learning to branch, 2022.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2026. URL `https://www.gurobi.com`.

Qingyu Han, Linxin Yang, Qian Chen, Xiang Zhou, Dong Zhang, Akang Wang, Ruoyu Sun, and Xiaodong Luo. A GNN-guided predict-and-search framework for mixed-integer linear programming. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

Zhinan Hou, Xingchen Li, Yankai Zhang, Tianxun Li, and Keyou You. Llm4branch: Large language model for discovering efficient branching policies of integer programs, 2026.

Taoan Huang, Aaron Ferber, Arman Zharmagambetov, Yuandong Tian, and Bistra Dilkina. Contrastive predict-and-search for mixed integer linear programs. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 19406–19424, Vienna, Austria, 2024.

Zeren Huang, Kerong Wang, Furui Liu, Hui-Ling Zhen, Weinan Zhang, Mingxuan Yuan, Jianye Hao, Yong Yu, and Jun Wang. Learning to select cuts for efficient mixed-integer programming, 2021.

Richard M. Karp. *Complexity of Computer Computations*. Plenum Press, 1972.

Elias B. Khalil, Pierre Le Bodic, Le Song, George Nemhauser, and Bistra Dilkina. Learning to branch in mixed integer programming. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 724–731, Phoenix, AZ, 2016.

Elias B. Khalil, Bistra Dilkina, George L. Nemhauser, Shabbir Ahmed, and Yufen Shao. Learning to run heuristics in tree search. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 659–666, Melbourne, Australia, 2017.

Yufei Kuang, Jie Wang, Haoyang Liu, Fangzhou Zhu, Xijun Li, Jia Zeng, Jianye Hao, Bin Li, and Feng Wu. Rethinking branching on exact combinatorial optimization solver: The first deep symbolic discovery framework. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

Kevin Leyton-Brown, Mark Pearson, and Yoav Shoham. Towards a universal test suite for combinatorial auction algorithms. In *Proceedings of the 2nd ACM Conference on Electronic Commerce*, pp. 66–76, Minneapolis, MN, 2000.

Xijun Li, Fangzhou Zhu, Hui-Ling Zhen, Weilin Luo, Meng Lu, Yimin Huang, Zhenan Fan, Zirui Zhou, Yufei Kuang, Zhihai Wang, Zijie Geng, Yang Li, Haoyang Liu, Zhiwu An, Muming Yang, Jianshu Li, Jie Wang, Junchi Yan, Defeng Sun, Tao Zhong, Yong Zhang, Jia Zeng, Mingxuan Yuan, Jianye Hao, Jun Yao, and Kun Mao. Machine learning insides optverse ai solver: Design principles and applications, 2024a.

Yixuan Li, Wanyuan Wang, Weiyi Xu, Yanchen Deng, and Weiwei Wu. Factor graph neural network meets max-sum: A real-time route planning algorithm for massive-scale trips. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1165–1173, Auckland, New Zealand, 2024b.

Jiacheng Lin, Meng Xu, Zhihua Xiong, and Huangang Wang. Cambranch: Contrastive learning with augmented milps for branching. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

Haotian Ling, Zhihai Wang, and Jie Wang. Learning to stop cut generation for efficient mixed-integer linear programming. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 20759–20767, Vancouver, Canada, 2024.

Haoyang Liu, Jie Wang, Zijie Geng, Xijun Li, Yuxuan Zong, Fangzhou Zhu, JianYe Hao, and Feng Wu. Apollo-milp: An alternating prediction-correction neural solving framework for mixed-integer linear programming. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, 2025.

Arjun M., Rijul Tandon, Agam Gupta, HARIPRASAD KODAMANA, and Manojkumar Ramteke. MIRACLE: Model-free imitation and reinforcement learning for adaptive cut-selection. In *Proceedings of the 14th International Conference on Learning Representations*, Rio de Janeiro, Brazil, 2026.

Kefan Ma, Liquan Xiao, Jianmin Zhang, and Tiejun Li. Accelerating an FPGA-based SAT solver by software and hardware co-design. *Chinese Journal of Electronics*, 28(5):953–961, 2019.

Vinod Nair, Sergey Bartunov, Felix Gimeno, Ingrid von Glehn, Pawel Lichocki, Ivan Lobov, Brendan O'Donoghue, Nicolas Sonnerat, Christian Tjandraatmadja, Pengming Wang, et al. Solving mixed integer programs using neural networks, 2020.

Max B. Paulus and Andreas Krause. Learning to dive in branch and bound. In *Advances in Neural Information Processing Systems 36*, pp. 34260–34277, New Orleans, LA, 2023.

Max B. Paulus, Giulia Zarpellon, Andreas Krause, Laurent Charlin, and Chris J. Maddison. Learning to cut by looking ahead: Cutting plane selection via imitation learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 17584–17600, Baltimore, MD, 2022.

Tianle Pu, Jianing Li, Yingying Gao, Shixuan Liu, Zijie Geng, Haoyang Liu, Chao Chen, and Changjun Fan. Coco-milp: Inter-variable contrastive and intra-constraint competitive milp solution prediction. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, pp. 24882–24890, Singapore, 2026.

Pol Puigdemont, Stratis Skoulakis, Grigorios Chrysos, and Volkan Cevher. Learning to remove cuts in integer linear programming. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 41235–41255, Vienna, Austria, 2024.

Lara Scavuzzo, Feng Chen, Didier Chetelat, Maxime Gasse, Andrea Lodi, Neil Yorke-Smith, and Karen Aardal. Learning to branch with tree mdps. In *Advances in Neural Information Processing Systems 35*, pp. 18514–18526, New Orleans, LA, 2022.

Yunhao Tang, Shipra Agrawal, and Yuri Faenza. Reinforcement learning for integer programming: Learning to cut. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9367–9376, Virtual, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, Long Beach, CA, 2017.

Jie Wang, Zhihai Wang, Xijun Li, Yufei Kuang, Zhihao Shi, Fangzhou Zhu, Mingxuan Yuan, Jia Zeng, Yongdong Zhang, and Feng Wu. Learning to cut via hierarchical sequence/set model for efficient mixed-integer programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9697–9713, 2024.

Ruobing Wang, Xin Li, and Mingzhong Wang. Milpnet: A multi-scale architecture with geometric feature sequence representations for advancing milp problems. In *Proceedings of the 14th International Conference on Learning Representations*, Rio de Janeiro, Brazil, 2026.

Zhihai Wang, Xijun Li, Jie Wang, Yufei Kuang, Mingxuan Yuan, Jia Zeng, Yongdong Zhang, and Feng Wu. Learning cut selection for mixed-integer linear programming via hierarchical sequence model. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

Taehyun Yoon. Confidence threshold neural diving, 2022.

Giulia Zarpellon, Jason Jo, Andrea Lodi, and Yoshua Bengio. Parameterizing branch-and-bound search trees to learn branching policies. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 3931–3939, Virtual, 2021.

Hao Zeng, Jiaqi Wang, Avirup Das, Junying He, Kunpeng Han, Haoyuan Hu, and Mingfei Sun. Effective generation of feasible solutions for integer programming via guided diffusion, 2024.

Changwen Zhang, Wenli Ouyang, Hao Yuan, Liming Gong, Yong Sun, Ziao Guo, Zhichen Dong, and Junchi Yan. Towards imitation learning to branch for mip: A hybrid reinforcement learning based sample augmentation approach. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

A Related Work

A.1 Learning alongside MILP Solvers

In recent years, machine learning has emerged as a promising paradigm for accelerating MILP solving (Li et al., 2024a), and existing studies generally fall into two categories. The first category learns heuristic policies for key components of branch-and-bound (B&B) solvers, including branching variable selection (Khalil et al., 2016; Gasse et al., 2019; Gupta et al., 2020; Zarpellon et al., 2021; Gupta et al., 2022; Scavuzzo et al., 2022; Lin et al., 2024; Zhang et al., 2024; Kuang et al., 2024), cutting plane selection (Tang et al., 2020; Wang et al., 2023; 2024; Huang et al., 2021; Paulus et al., 2022; Ling et al., 2024; Puigdemont et al., 2024), and scheduling of primal heuristics Khalil et al. (2017); Chmiela et al. (2021). These methods typically leverage imitation learning or reinforcement learning to learn effective heuristic policies from historical solving trajectories or expert policies.

Among all components, branching and cut selection have received the most attention due to their significant impact on solver performance. For branching, Khalil et al. (2016) pioneered the imitation of strong branching decisions via a learned ranking function, and more recently, Hou et al. (2026) leveraged large language models to automatically discover branching policies. For cut selection, Tang et al. (2020) proposed to learn adaptive selection policies through reinforcement learning, while M. et al. (2026) further combined imitation learning and reinforcement learning to obtain memory-efficient policies. Collectively, these studies demonstrate that learning effective policies for individual solver components can substantially enhance the performance of modern MILP solvers.

A.2 End-to-End Prediction for MILP

The second category employs machine learning models such as graph neural networks (GNNs) to directly predict high-quality solutions for MILP instances (Ding et al., 2020; Zeng et al., 2024; Geng et al., 2025). Research in this category advances along two complementary aspects. The first aspect concerns how predictions are exploited to reduce the original problem (Nair et al., 2020; Han et al., 2023; Liu et al., 2025). Nair et al. (2020) proposed to directly fix high-confidence variables to their predicted values, yielding a reduced sub-problem, which was subsequently refined in follow-up studies (Yoon, 2022; Paulus & Krause, 2023). Han et al. (2023) and Huang et al. (2024) relaxed such hard fixing by searching within a trust region centered at the predicted partial assignment, and Liu et al. (2025) further extended this paradigm by alternating between prediction and solver-based correction.

The second aspect focuses on improving the prediction capacity (Pu et al., 2026; Wang et al., 2026). Pu et al. (2026) enhanced prediction quality by explicitly modeling inter-variable correlations, while Wang et al. (2026) replaced the conventional bipartite graph representation with geometric feature sequences and performed prediction via a multi-scale attention architecture. Together, these efforts have established learning-based solution prediction as an effective paradigm for accelerating MILP solving.

B Graph Representation Details

The bipartite graph representation used in this paper is based on the representation adopted by Predict-and-Search (Han et al., 2023). To support hierarchical decoding, we further introduce three additional variable features: target indicator, previous prediction, and previous confidence. We list the graph features in Table 6.

Table 6: The variable features, constraint features, and edge features used for the graph representation. Newly introduced features are shown in bold.

Index	Variable Feature Name	Description
0	Objective	Normalized objective coefficient
1	Variable coefficient	Average variable coefficient in all constraints
2	Variable degree	Degree of the variable node in the bipartite graph representation
3	Maximum variable coefficient	Maximum variable coefficient in all constraints
4	Minimum variable coefficient	Minimum variable coefficient in all constraints
5	Variable type	Whether the variable is an integer variable or not
6	**Target indicator**	**Whether the variable belongs to the current prediction hierarchy**
7	**Previous prediction**	**Previously predicted value of the variable**
8	**Previous confidence**	**Confidence score of the predicted value**
9–24	Position embedding	Binary encoding of the order of appearance for each variable among all variables
Index	**Constraint Feature Name**	**Description**
0	Constraint coefficient	Average of all coefficients in the constraint
1	Constraint degree	Degree of constraint nodes
2	Bias	Normalized right-hand side of the constraint
3	Sense	The sense of the constraint
Index	**Edge Feature Name**	**Description**
0	Coefficient	Constraint coefficient

C VARIABLE COUPLING GRAPH DETAILS

Expected Violation The *expected violation* measures how strongly a variable pair is restricted by a constraint. Violation of the variable pair (z_i, z_j) in constraint c is defined as

$$\Delta\left(\phi_{ij}^{(c)}\right) = \begin{cases} \max(b_c - \phi_{ij}^{(c)}, 0), & \circ_c \text{ is } \geq, \\ \max(\phi_{ij}^{(c)} - b_c, 0), & \circ_c \text{ is } \leq, \\ \left|\phi_{ij}^{(c)} - b_c\right|, & \circ_c \text{ is } = . \end{cases}$$

where $\phi_{ij}^{(c)} = a_{ci} z_i + a_{cj} z_j$, with a_{ci} and a_{cj} denoting the coefficients of variables z_i and z_j in constraint c, respectively.

The expected violation $P_{ij}^{(c)} = \mathbb{E}\left[\Delta\left(\phi_{ij}^{(c)}\right)\right]$ is computed as follows. For a binary–binary pair (z_i, z_j), we assume $z_i, z_j \sim \mathrm{Uniform}(\{0,1\})$, and the expected violation is computed as

$$P_{ij}^{(c)} = \frac{1}{4} \sum_{z_i, z_j \in \{0,1\}} \Delta\left(\phi_{ij}^{(c)}\right).$$

For a binary–continuous pair (z_i, z_j), we assume $z_i \sim \mathrm{Uniform}(\{0,1\})$ and $z_j \sim \mathrm{Uniform}(l_j, u_j)$, where l_j and u_j denote the lower and upper bounds of z_j. The expected violation is computed as

$$P_{ij}^{(c)} = \frac{1}{2} \sum_{z_i \in \{0,1\}} \frac{1}{u_j - l_j} \int_{l_j}^{u_j} \Delta\left(\phi_{ij}^{(c)}\right) dz_j.$$

To eliminate scale differences across constraints and measure the relative strength of a variable pair within the constraint, we normalize the expected violation as

$$\widehat{P}_{ij}^{(c)} = \frac{P_{ij}^{(c)}}{\frac{1}{|E_c|}\sum_{(u,v)\in E_c} P_{uv}^{(c)}},$$

where E_c denotes the set of retained variable pairs in constraint c.

Coefficient Importance The *coefficient importance* characterizes the relative contribution of a variable pair within a constraint. For each variable z_i, we define a scale factor s_i according to its variable type, where s_i is set to 1 for binary variables and to $u_i - l_i$ for continuous variables. Then we compute the coefficient importance $r_{ij}^{(c)}$ of the variable pair and normalize it within the constraint to obtain $\widehat{r}_{ij}^{(c)}$, which represents the relative importance of a variable pair within the constraint:

$$r_{ij}^{(c)} = \frac{|a_{ci}|s_i + |a_{cj}|s_j}{2},$$
$$\widehat{r}_{ij}^{(c)} = \frac{r_{ij}^{(c)}}{\frac{1}{|E_c|}\sum_{(u,v)\in E_c} r_{uv}^{(c)}}.$$

Edge Weight Aggregation. Finally, the local coupling weight of a pair in constraint c is computed as $w_{ij}^{(c)} = \widehat{P}_{ij}^{(c)}\widehat{r}_{ij}^{(c)}$, and the final edge weight W_{ij} between variables z_i and z_j is obtained by summing the local weights over all constraints containing both variables.

D Benchmark Details

D.1 Subset of MIPLIB

We use the IIS dataset, which is a subset of MIPLIB (Gleixner et al., 2021) constructed by Liu et al. (2025), to evaluate the solvers' ability to handle challenging real-world instances. According to Liu et al. (2025), the instances in the dataset are selected based on their similarity, which is measured using the 100 human-designed features (Gleixner et al., 2021). Instances with presolving times exceeding 300 seconds or those that exceed GPU memory limits during the inference process are discarded. The IIS dataset contains eleven instances, including eight training instances and three testing instances (ramos3, scpj4scip, and scpl4). Table 7 reports detailed statistics of the IIS dataset.

Table 7: Statistical information of the instances in the IIS dataset. All instances contain only binary variables.

Instance	Constraint Number	Variable Number	Binary Variables
ex1010-pi	1468	25200	25200
fast0507	507	63009	63009
glass-sc	6119	214	214
iis-glass-cov	5375	214	214
iis-hc-cov	9727	297	297
ramos3	2187	2187	2187
scpj4scip	1000	99947	99947
scpk4	2000	100000	100000
scpl4	2000	200000	200000
seymour	4944	1372	1372
v150d30-2hopcds	7822	150	150

E Implementation Details

E.1 SHSP Predictor Details

The predictor used in SHSP is based on the graph neural network architecture adopted in previous learning-based MILP approaches (Han et al., 2023; Liu et al., 2025). To support the hierarchical

decoding process in SHSP, we introduce a learnable step embedding, which enables the network to adapt its prediction behavior across different decoding steps. Specifically, the initial representation of each variable node is computed as

$$h_{v_i}^{(0,k)} = f_s\left(\left[f_v\left(x_i\right); \mathrm{Emb}(k)\right]\right).$$

Here, v_i denotes the i-th variable node, k is the current decoding step, and x_i is the feature vector of v_i. The function $f_v(\cdot)$ denotes the variable embedding network, $\mathrm{Emb}(k)$ is the learnable embedding corresponding to decoding step k, and $f_s(\cdot)$ fuses the variable and step embeddings. The resulting representation $h_{v_i}^{(0,k)}$ is then used as the input to the same bipartite message-passing architecture adopted in previous approaches (Han et al., 2023; Liu et al., 2025).

E.2 Training Details

During predictor training, we set the initial learning rate to be 0.001 and the number of training epochs to 500. To collect the training data, we run a single-threaded Gurobi (Gurobi Optimization, LLC, 2026) on each training and validation instance for 3,600 seconds and record the best 50 solutions found during the run.

For SHSP training, we incorporate a *teacher forcing* strategy that employs ground-truth assignments as conditioning information in the early stage of training, and the teacher forcing ratio ρ_t is progressively decayed as training proceeds. Specifically, with probability ρ_t, we use the ground-truth value as the history input for the next step; otherwise, we use the model prediction. The teacher forcing ratio ρ_t is linearly decayed from ρ_{start} to ρ_{end} during the first T_{decay} epochs:

$$\rho_t = \rho_{\mathrm{start}} + \min\left(1, \frac{t}{T_{\mathrm{decay}}}\right)\left(\rho_{\mathrm{end}} - \rho_{\mathrm{start}}\right),$$

where t is the current epoch. During validation and inference, we disable teacher forcing and always use the model predictions to update the conditioning context. In experiments, we set $\rho_{\mathrm{start}} = 1.0$, $\rho_{\mathrm{end}} = 0$, and $T_{\mathrm{decay}} = 50$.

E.3 Structure-Aware Fixing Strategies Details

The proposed structure-aware variable fixing strategy jointly considers the variable coupling structure and the prediction confidences. We first identify variables whose variable coupling scores exceed a prescribed threshold as high-coupling variables. The identified variables are subsequently screened according to their prediction confidences, where variables with $p_i \geq \theta_1$ are eligible for fixing to one, and those with $p_i \leq \theta_0$ are eligible for fixing to zero.

We further cap the number of fixed variables. The variables selected above are ranked by their prediction probabilities, and at most k_1 and k_0 of them are retained for fixing to one and to zero, respectively. Note that k_1 and k_0 serve only as upper bounds. If the thresholds yield fewer variables, no additional ones are fixed to reach these bounds. To ensure a fair comparison, the settings of k_0 and k_1 are kept consistent with those of existing methods.

E.4 Integration with Downstream Learning-based Search Methods

We replace the original GNN predictor and the variable fixing strategy in each framework with our SHSP predictor and structure-aware variable fixing strategy, forming **ND+SHSP**, **PaS+SHSP**, and **Apollo+SHSP**, respectively.

For ND+SHSP and PaS+SHSP, the variable coupling graph is constructed once before neural network inference for each instance. The graph construction time is included in the solver time budget.

For Apollo+SHSP, the variable coupling graph is constructed before the first prediction and is subsequently updated for each reduced subproblem before prediction. Similarly, the time required for graph construction and updates is deducted from the corresponding solver time budget.

E.5 Inference Details

We conducted all the experiments on a single machine with an AMD EPYC 7513 32-Core Processor and NVIDIA GeForce RTX 4090 GPUs. The inference settings are described as follows.

For the baselines, the hyperparameters (k_0, k_1) for ND and (k_0, k_1, Δ) for PaS are summarized in Table 8. For Apollo-MILP, the number of iterations is set to 4. The hyperparameter settings $(k_0^{(i)}, k_1^{(i)}, \Delta^{(i)})$ for each iteration are summarized in Table 9.

Table 8: Hyperparameter settings for PaS and ND on different benchmark datasets.

Dataset	PaS (k_0, k_1, Δ)	ND (k_0, k_1)
CA	$(600, 0, 20)$	$(600, 0)$
WA	$(0, 500, 10)$	$(0, 500)$
IP	$(400, 5, 10)$	$(400, 5)$
SC	$(2000, 0, 100)$	$(2000, 0)$

Table 9: Hyperparameter configuration of $(k_0^{(i)}, k_1^{(i)}, \Delta^{(i)})$ in Apollo-MILP.

	CA	WA	IP	SC
Iteration 1	$(400, 0, 60)$	$(20, 200, 100)$	$(100, 20, 50)$	$(1000, 0, 200)$
Iteration 2	$(200, 0, 30)$	$(10, 100, 50)$	$(40, 15, 20)$	$(500, 0, 100)$
Iteration 3	$(100, 0, 15)$	$(10, 5, 5)$	$(20, 15, 10)$	$(250, 0, 50)$
Iteration 4	$(50, 0, 10)$	$(1, 10, 5)$	$(5, 50, 30)$	$(10, 0, 5)$

For our method, we use the same hyperparameter settings as the corresponding baseline framework. The number of decoding steps is set to 2 for all benchmark datasets except WA, for which it is set to 4. The additional hyperparameter settings $(\theta_0, \theta_1, \eta)$ are reported in Table 10, where η controls the VCS threshold. Specifically, let $z_{(1)}, \ldots, z_{(n)}$ denote the variables sorted in descending order of VCS. We use the VCS of $z_{(\lceil \eta n \rceil)}$ as the threshold.

Table 10: The additional hyperparameter settings $(\theta_0, \theta_1, \eta)$ for ND+SHSP, PaS+SHSP and Apollo+SHSP on different benchmark datasets.

Dataset	ND+SHSP	PaS+SHSP	Apollo+SHSP
CA	$(0.10, 0.90, 0.5)$	$(0.10, 0.90, 0.5)$	$(0.04, 0.90, 0.7)$
WA	$(0.10, 0.90, 0.7)$	$(0.10, 0.90, 0.7)$	$(0.10, 0.95, 0.8)$
IP	$(0.05, 0.95, 0.7)$	$(0.10, 0.90, 0.7)$	$(0.10, 0.50, 0.5)$
SC	$(0.10, 0.90, 0.5)$	$(0.10, 0.90, 0.5)$	$(0.10, 0.90, 0.7)$

F Additional Experimental Results

F.1 More Ablation Study Results

Confidence-Based Mask and Repair Mechanism We perform an ablation study to evaluate the proposed mask-and-repair mechanism. Specifically, we compare three variants: None (without mask or repair), Mask (mask only), and Mask+Repair (mask and repair) on the four benchmark datasets (CA, WA, IP, and SC). The results are shown in Table 11. The observations are consistent with those discussed in the main paper: the two components are complementary and jointly contribute to the overall performance.

Hierarchical Solution Prediction Predictor To evaluate the effectiveness of the proposed hierarchical solution prediction predictor, we compare it with the original GNN predictor under the same structure-aware fixing strategy on the four benchmark datasets (CA, WA, IP, and SC). The results are shown in Table 12, which suggests that our predictor generates more accurate solution predictions, thereby enabling more effective variable fixing and problem reduction.

Fixing Strategies To evaluate the effectiveness of the proposed structure-aware fixing strategy, we compare it with the original fixing strategy while using the same predictor. The results on the four

Table 11: Detailed ablation study of the proposed confidence-based mask-and-repair strategy. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. Bold values denote the best results, including ties.

Method	CA ↑(BKS:98489.24)		WA ↓(BKS:706.86)		IP ↓(BKS:11.69)		SC ↓(BKS:123.30)	
	Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}
PaS+SHSP (None)	98380.92	108.32	707.06	0.20	12.12	0.43	123.47	0.17
PaS+SHSP (Mask)	98416.55	72.69	707.05	0.19	12.17	0.48	123.53	0.23
PaS+SHSP (Mask+Repair)	**98487.72**	**1.52**	**706.99**	**0.13**	**11.97**	**0.28**	**123.41**	**0.11**
Apollo+SHSP (None)	98385.20	104.04	707.10	0.24	11.78	0.09	123.42	0.12
Apollo+SHSP (Mask)	98344.07	145.17	**707.01**	**0.15**	11.80	0.11	123.39	0.09
Apollo+SHSP (Mask+Repair)	**98489.24**	**0.00**	**707.01**	**0.15**	**11.71**	**0.02**	**123.37**	**0.07**

Table 12: Detailed ablation study of the proposed hierarchical solution prediction predictor and structure-aware fixing strategy. ↑ indicates that higher values are better, while ↓ indicates that lower values are better. Bold values denote the best results, including ties.

Method	CA ↑(BKS:98489.24)		WA ↓(BKS:706.86)		IP ↓(BKS:11.69)		SC ↓(BKS:123.30)	
	Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}	Obj	Gap_{abs}
PaS (structure-aware fixing)	97184.33	1304.91	707.08	0.22	12.12	0.43	123.49	0.19
PaS+SHSP (original fixing)	98061.51	427.73	707.01	0.15	12.34	0.65	123.53	0.23
PaS+SHSP (structure-aware fixing)	**98487.72**	**1.52**	**706.99**	**0.13**	**11.97**	**0.28**	**123.41**	**0.11**
Apollo (structure-aware fixing)	98015.32	473.92	707.09	0.23	11.87	0.18	123.40	0.10
Apollo+SHSP (original fixing)	98070.43	418.81	**707.01**	**0.15**	12.08	0.39	**123.37**	**0.07**
Apollo+SHSP (structure-aware fixing)	**98489.24**	**0.00**	**707.01**	**0.15**	**11.71**	**0.02**	**123.37**	**0.07**

benchmark datasets (CA, WA, IP, and SC) are shown in Table 12. The observations are consistent with those discussed in the main paper: prioritizing structurally important variables leads to more effective search space reduction, allowing the solver to explore the remaining search space more efficiently and achieve better performance.

F.2 Results on MIPLIB

To further demonstrate the applicability of our method, we conduct experiments on instances from the challenging real-world dataset MIPLIB (Gleixner et al., 2021). However, applying ML-based solvers directly to the entire dataset can be difficult because of the heterogeneous nature of the instances in MIPLIB. Therefore, following Liu et al. (2025), we focus on a subset of MIPLIB that contains similar instances. More information on the selected MILP subset, referred to as IIS, is provided in Appendix D.1. We report the solving performance of the solvers in Table 13 and Table 14, where our methods outperform their corresponding baselines, showcasing the potential for real-world applications.

Table 13: Results on the IIS dataset, a subset of MIPLIB used by Liu et al. (2025). All learning-based methods use Gurobi as the downstream solver, with a solving time limit of 3,600 seconds.

Method	Obj↓	Gap_{abs} ↓
Gurobi	211.00	20.00
PaS	209.00	18.00
PaS+SHSP	**206.67**	**15.67**
Apollo	214.33	23.33
Apollo+SHSP	**208.67**	**17.67**

F.3 Runtime Results

We report network inference time and graph construction time in Table 15. The inference time of the SHSP predictor is only slightly higher than that of the original GNN predictor, which remains

Table 14: The best objectives found on each test instance in IIS. BKS denotes the best known objective values reported by MIPLIB (`https://miplib.zib.de/index.html`).

Instance	BKS	Gurobi	PaS	Apollo	PaS+SHSP	Apollo+SHSP
ramos3	186.00	228.00	227.00	242.00	**221.00**	226.00
scpj4scip	128.00	133.00	131.00	132.00	**130.00**	131.00
scpl4	259.00	272.00	269.00	269.00	269.00	269.00

negligible compared with the overall solver runtime. In addition, the graph construction time ranges from 0.26s to 6.77s per instance, which is also much shorter than the solver runtime.

Table 15: Comparison of the average per-instance inference time and graph construction time (in seconds) on the four benchmark datasets.

Dataset	GNN Predictor	SHSP Predictor	Graph construction time
CA	0.0050s	0.0146s	0.82s
WA	0.0439s	0.2197s	1.75s
IP	0.0043s	0.0131s	0.26s
SC	0.0658s	0.1953s	6.77s