

Retrieval-Augmented Fine-Tuning With Preference Optimization For Visual Program Generation

Deokhyung Kang^{*1}, Jeonghun Cho^{*1}, Yejin Jeon¹, Sunbin Jang³,
Minsub Lee³, Jawoon Cho⁴, Gary Geunbae Lee^{1,2}

¹Graduate School of Artificial Intelligence, POSTECH,

²Department of Computer Science and Engineering, POSTECH,

³Hyundai Mobis, ⁴T&I Company

{deokhk, jeonghuncho, jeonyj0612, gblee}@postech.ac.kr

{soonbin.Jang, perfect}@mobis.com, jwcho@tnicompany.com

Abstract

Visual programming languages (VPLs) allow users to create programs through graphical interfaces, which results in easier accessibility and their widespread usage in various domains. To further enhance this accessibility, recent research has focused on generating VPL code from user instructions using large language models (LLMs). Specifically, by employing prompting-based methods, these studies have shown promising results. Nevertheless, such approaches can be less effective for industrial VPLs such as Ladder Diagram (LD). LD is a pivotal language used in industrial automation processes and involves extensive domain-specific configurations, which are difficult to capture in a single prompt. In this work, we demonstrate that training-based methods outperform prompting-based methods for LD generation accuracy, even with smaller backbone models. Building on these findings, we propose a two-stage training strategy to further enhance VPL generation. First, we employ retrieval-augmented fine-tuning to leverage the repetitive use of subroutines commonly seen in industrial VPLs. Second, we apply direct preference optimization (DPO) to further guide the model toward accurate outputs, using systematically generated preference pairs through graph editing operations. Extensive experiments on real-world LD data demonstrate that our approach improves program-level accuracy by over 10% compared to supervised fine-tuning, which highlights its potential to advance industrial automation.

1 Introduction

Recent advances in large language models (LLMs) have significantly improved their capabilities in code generation. Moreover, LLMs such as GPT-4 (Achiam et al., 2023), StarCoder (Li et al., 2023), and DeepSeek-Coder (Guo et al., 2024) are able

to automate large parts of programming tasks and substantially improve programmers' efficiency. Yet, despite these achievements, most of the previous research has focused on text-based programming languages (TPLs) such as Python or Java, which leaves visual programming languages (VPLs) relatively unexplored.

VPLs typically represent programs as node graphs (Figure 1), which allow users with limited programming backgrounds to create and modify programs by graphical manipulation (DeLozier and Shey, 2023). As such, due to their easier accessibility, VPLs have been widely adopted across various domains, from Ladder Diagram (LD) in industrial control systems (IEC, 2013) to Unreal Engine's Blueprints in game development (Games, 2010). Although VPLs lower the barrier to programming, creating such visual programs from scratch can still be cumbersome. Consequently, recent studies (Zhang et al., 2024b; Xue et al., 2024; Zhou et al., 2025) have explored generating visual programs in text formats (e.g., JSON) from user instructions using prompting-based approaches.

While these studies have shown promising results in VPL generation, the sole reliance on prompting-based methods can be less effective for industrial VPLs like LD. This is because these languages are widely used in industrial automation, which follow domain-specific configurations (e.g., address mapping) that vary drastically by environments (Alphonsus and Abdullah, 2016). Given the vast number of these configurations, it is challenging to include all such information in a single prompt. In contrast, training-based approaches can implicitly learn these configurations during fine-tuning. Therefore, we argue that training-based methods are necessary for this setting.

To validate this, we select LD as a test language and compare supervised fine-tuning (SFT) with retrieval-augmented generation (RAG), where the latter is a commonly used prompting-based method

^{*}Equally contributed

for VPL generation (Zhang et al., 2024b; Xue et al., 2024). Experimental results show that SFT outperforms RAG even with a smaller LLM backbone, which highlights the advantage of training-based approaches for industrial VPL generation.

Motivated by these findings, we propose a two-stage training approach to enhance VPL generation. First, we adopt retrieval-augmented fine-tuning, as VPL often reuses *subroutines* in similar contexts (Terra-Neves et al., 2021). During both training and inference, we retrieve similar examples from a corpus and incorporate them into the model input to improve generation quality. Second, we apply offline direct preference optimization (DPO) (Rafailov et al., 2023) to further guide the model toward accurate outputs. Since VPLs are represented as node-based graphs, we construct preference pairs by systematically applying graph editing operations to transform ground-truth (preferred) VPL code into unpreferred variants. By learning from these pairs, the model captures the preference relationship between the preferred and unpreferred code, thereby enhancing its fine-grained understanding of VPL’s structural properties.

To evaluate our approach, we collect LD data from an actual production environment and conduct extensive evaluations across various text formats and models. Results demonstrate a consistent improvement in generation performance. Notably, our approach improves exact-match accuracy at the program level by more than 10% compared to SFT, significantly reducing manual inspection efforts and thus accelerating PLC development. Further analyses and ablation studies substantiate the advantages of our methodology.

Our contributions are as follows: (1) We present a pioneering study on training-based VPL generation and demonstrate its effectiveness for industrial VPL. (2) We propose a two-stage training strategy that combines retrieval-based fine-tuning and preference optimization via graph editing, which yields significant performance improvements. (3) To the best of our knowledge, this is the first work to effectively generate Ladder Diagrams using LLMs, underscoring its potential to enhance industrial automation.

2 Is Training-Based Approach Effective?

In this section, we investigate the effectiveness of the training-based approach in comparison to the prompting-based approach for industrial VPL gen-

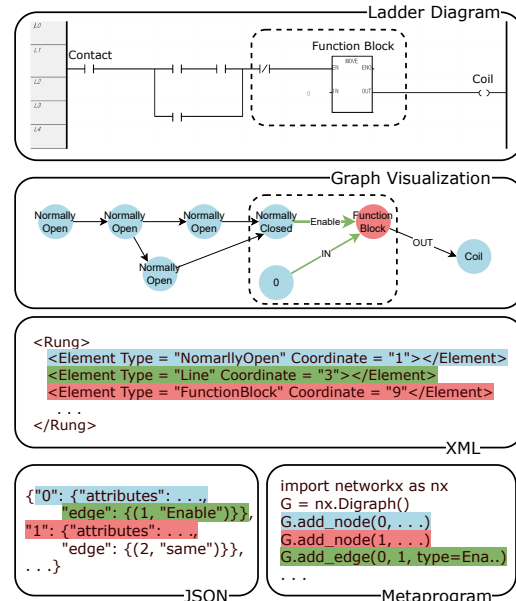


Figure 1: The topmost subfigure shows a single rung of LD and its corresponding visualized graph. The bottom displays XML tags exportable from a Ladder Diagram IDE, along with JSON and Metaprogram representations that capture structural relationships in the graph.

eration. We introduce Ladder Diagram (LD) as our test language (§2.1), then describe text format conversion in §2.2 to explore results across different text formats. The evaluation metrics used for comparison are detailed in §2.3, and the corresponding results are presented in §2.4.

2.1 Ladder Diagram

Programmable Logic Controller (PLC) (Erickson, 1996) controls physical devices such as sensors and actuators in industrial automation systems. For example, in a conveyor system, PLC logic enables a robotic arm to detect products via sensors and control its joints to transfer items between conveyors. This logic is typically implemented using Ladder Diagram (LD) (Malik, 2024), which consists of multiple rungs, each executing a specific task (e.g., operating a motor). Figure 1 illustrates an LD with a single rung.

Each rung is further composed of contacts, coils, and function blocks and can be represented as a node-based graph, where each visual element is a node. These individual elements serve a distinct role in the logic sequence; contacts control the power flow, coils activate machinery, and function blocks (e.g., timers, counters) provide advanced

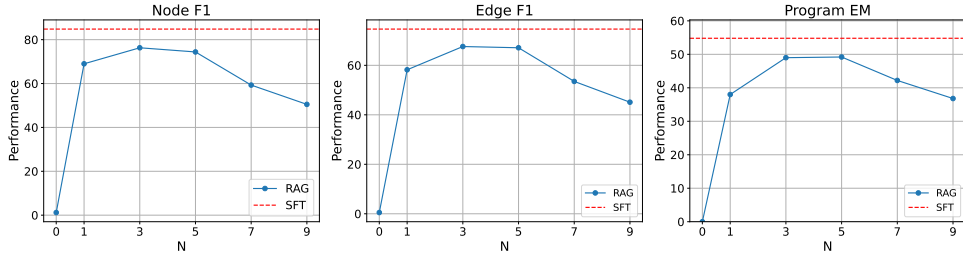


Figure 2: Performance comparison between SFT and RAG, where RAG uses a larger LLM. N represents the number of retrieved examples in RAG, and SFT’s performance is represented by a red dotted line. We use XML as the text format.

Format	Method	Node F1 \uparrow	Edge F1 \uparrow	Node EM \uparrow	Edge EM \uparrow	Program EM \uparrow
XML	RAG	74.4	67.1	49.6	49.2	49.2
	SFT	84.8	74.7	55.2	54.8	54.8
JSON	RAG	72.6	65.7	51.6	50.8	50.8
	SFT	85.1	74.9	53.6	52.6	52.6
Metaprogram	RAG	80.0	71.9	52.2	51.8	51.8
	SFT	86.3	75.7	55.2	54.0	54.0

Table 1: Performance across different text formats. The highest values in each format are **in bold**.

control functions. They are connected to the PLC and mapped to I/O addresses that vary with the environment and hardware setup. This domain-specific address mapping enables customized control, as each rung manages unique I/O settings tailored to its setup.

2.2 Text Format Conversion

As LLMs cannot generate complex visual elements directly, prior studies (Xue et al., 2024; Zhang et al., 2024b) generate VPL in text formats. As each format has unique characteristics, we convert VPL into various text formats to conduct comprehensive experiments. We consider three standard text formats for VPL: (1) XML, which represents visual elements sequentially; (2) JSON, which explicitly captures relationships between elements; (3) Metaprogram (code), which encodes VPL using a code-based syntax.

Specifically, Figure 1 illustrates that XML format represents LD as a list of visual elements, including contacts, lines, and function blocks. Each visual element corresponds to an `<Element>` tag containing its element type, coordinates, and other attributes. Since XML does not explicitly define relationships between elements, we extract these relationships using rule-based methods and convert them into a graph. We iterate through the visual elements in coordinate-based order, adding all elements except lines as nodes to the graph. We

implement this graph using NetworkX (Hagberg et al., 2008), which is a widely used Python library for graph representation. We assign Node IDs starting from 0, increasing sequentially in coordinate order. To determine edges, we analyze node coordinates and line positions. Using this information, we construct a directed acyclic graph (DAG).

Using this graph representation of LD, we represent LD in both JSON and metaprogram formats. In the JSON format, we represent the LD as a dictionary; each node ID in the graph is a key, and its attributes and outgoing node IDs form the values. For the metaprogram (code) format, we represent the LD in Python syntax using NetworkX. We traverse the graph starting from the smallest node ID, visiting each node and its successors. Upon first visiting a node, we append a `G.add_node(...)` statement with its attributes to the code. Similarly, we append a `G.add_edge(...)` statement to the code for each neighbor relationship. This process continues until all nodes from the graph are visited. The generated code can be executed to reconstruct the original graph.

2.3 Evaluation Metrics

While text-based programming languages such as Python are typically assessed using unit tests (Chen et al., 2021; Austin et al., 2021; Chen et al., 2023), VPLs often depend on external simulators (Ray, 2017; Ren et al., 2024), making consistent auto-

mated evaluation challenging across varying environments. To address this, we propose a graph-based automatic evaluation method. We first transform LD programs into NetworkX graphs as described in §2.2. For each LD program, we obtain a ground-truth graph $G = (V, E)$ and a predicted graph $\hat{G} = (\hat{V}, \hat{E})$, which are then compared from two perspectives: partial correctness and exact match.

Partial correctness is measured using F1 scores:

$$\text{Node F1} = \frac{2|V \cap \hat{V}|}{|V| + |\hat{V}|}, \quad \text{Edge F1} = \frac{2|E \cap \hat{E}|}{|E| + |\hat{E}|}$$

Here, true positives, false positives, and false negatives are computed based on set overlaps. A node (or edge) is considered a match only if all required attributes (e.g., names, types) are exactly equal.

Exact Match (EM) scores assess complete correctness: Node EM is defined as $1\{\hat{V} = V\}$, Edge EM as $1\{\hat{E} = E\}$, and Program EM as $1\{\hat{V} = V \text{ and } \hat{E} = E\}$. Program EM equals 1 only when both the node and edge sets exactly match those of the ground-truth, meaning that the predicted program is identical to it. For a concrete example of metric computation, refer to Appendix A.

2.4 Results

Training-based method outperforms prompting-based method To evaluate the effectiveness of the training-based approach compared to the prompting-based approach for industrial VPL generation, we conduct a comparison between the supervised fine-tuning (SFT) method and Retrieval-Augmented Generation (RAG). Among prompting-based methods, we select RAG because this approach has been widely used in prior studies on VPL generation (Xue et al., 2024; Zhang et al., 2024b), and retrieval can provide examples that reflect the domain-specific configurations relevant to generation. SFT uses Llama3.1-8B-Instruct model as the backbone, while RAG uses Llama-3.1-70B-Instruct¹. The RAG approach utilizes BM25 (Robertson et al., 2009) to retrieve N similar prompt-code pairs from the SFT training dataset. We then append them to the input to generate code for the test prompt². To assess the model’s

¹Due to computational constraints, we employed the AWQ-quantized (Lin et al., 2024) model: hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4

²We conducted our experiments using the dataset described in §4.

internal capabilities, we also consider the $N = 0$ setting (i.e., zero-shot). However, since the prompt template does not contain a concrete example of an LD in text format, most of the generated LDs are ill-formed. Therefore, we provide a fixed prompt-code pair even for $N = 0$ to inform the model of the expected format.

As shown in Figure 2, SFT consistently outperforms RAG across different numbers of retrieved examples (1, 3, 5, 7, and 9), despite its smaller backbone. Although RAG’s performance improves as N increases, it peaks at $N = 5$ and degrades thereafter, suggesting that learning domain-specific configurations in industrial VPL from retrieved examples alone is challenging. For the zero-shot setting, the model performs poorly, with a Node F1 score of just 1.2. This is likely because domain-specific LD configurations, which vary significantly between environments, are not present in the LLM’s pre-training data.

Overall, these findings demonstrate that SFT is a more effective approach in this setting. We observe similar trends with a different LLM (Appendix F), and the prompt template for the RAG model is provided in Appendix E.

Training-based method excels across text formats

To investigate whether the prior results are consistent across different text formats, we extend our experiments to include JSON and the metaprogram formats. We compare SFT with RAG, where performance is reported based on the number of retrieved examples that achieves the highest Program EM for each text format, as the optimal number of examples can vary across formats.

From the results in Table 1, we can observe the following: (1) While SFT shows stable performance across different formats, RAG exhibits performance differences in Node/Edge F1, with the metaprogram-based format showing better performance. This aligns with prior results on prompting-based approaches from Xue et al. (2024). (2) SFT outperforms RAG across all formats. These results further validate the effectiveness of the training-based approach.

3 Proposed Methodology

In this section, we present a two-stage training strategy to further improve the accuracy of VPL generation. Figure 3 illustrates an overview of our proposed methodology.

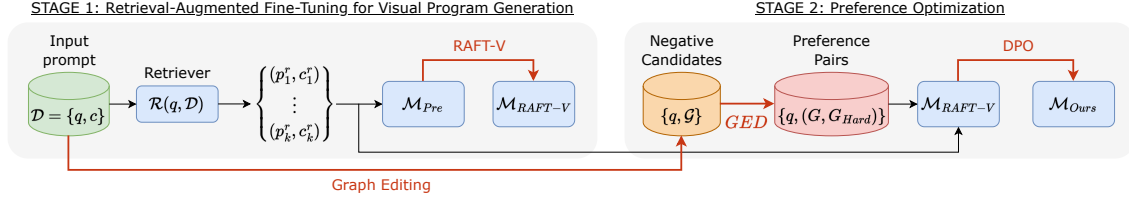


Figure 3: An overview of the two-stage training method. (1) RAFT-V: An off-the-shelf retriever is utilized for relevant prompt augmentation, and training is conducted with cross-entropy loss. (2) Preference Optimization: Preference learning leverages graph-edited preference pairs, with retrieved prompt-code pairs as additional input.

3.1 Stage 1: Retrieval-augmented Fine-Tuning for Visual Program Generation (RAFT-V)

Visual programming languages often contain recurring subroutines and modules that are reused in similar contexts (Terra-Neves et al., 2021). To leverage this property, we retrieve relevant examples from a training data pool and use them to guide the generation process. Inspired by Zhang et al. (2024a), which incorporates related documents as context during fine-tuning to improve domain-specific question answering, we extend this idea to the generation of VPL code. Specifically, we retrieve relevant prompt-code pairs as context to enhance the model’s performance. Let \mathcal{D} be a training dataset consisting of N prompt-code pairs, where p is a *prompt* and c is a corresponding VPL *code*. Given a new input prompt q , we aim to generate the target code c_q . We augment the input to the model with additional prompt-code pairs that are most similar to q . This process is driven by a *retriever* \mathcal{R} , which identifies similar prompts from \mathcal{D} . \mathcal{R} ranks all prompts in \mathcal{D} based on their similarity to q . We then select the top- k most similar prompts:

$$\{(p_i^r, c_i^r)\}_{i=1}^k = \mathcal{R}(q, \mathcal{D}) \quad (1)$$

where p_i^r and c_i^r denote the retrieved prompt and code, respectively. We then concatenate the original prompt q with each retrieved pair (p_i^r, c_i^r) to form the augmented input q_a for the model:

$$q_a = (q, [(p_1^r, c_1^r), (p_2^r, c_2^r), \dots, (p_k^r, c_k^r)]) \quad (2)$$

During fine-tuning, the model takes the q_a and learns to generate the target code c_q from q_a via SFT. At inference, the trained model predicts the code \hat{c}_q from q_a .

3.2 Stage 2: Preference Optimization

Preference optimization techniques, such as DPO (Rafailov et al., 2023), are widely used as

Algorithm 1 Graph Editing

```

1: Input: Random seed set  $S$  with  $|S| = s$ , Deletion ratio  $\tau$ , Graph  $G = (V, E)$  with  $|V(G)| = n$  nodes and  $|E(G)| = m$  edges
2: // Initialize the graph
3:  $G' \leftarrow G, \mathcal{G} \leftarrow \emptyset$ ;
4: // Compute number of nodes to delete
5:  $k \leftarrow \lfloor \tau \cdot |V(G')| \rfloor$ ;
6: for each  $i$  in  $S$  do
7:   if  $k > 0$  then
8:     // Randomly select and reconnect nodes
9:      $\mathcal{X} \leftarrow$  random  $k$  nodes from  $G'$ ;
10:    Let  $\mathcal{W} \rightarrow \mathcal{X} \rightarrow \mathcal{Y}$  be a path in  $G'$ ;
11:    Connect  $\mathcal{W} \rightarrow \mathcal{Y}$ , preserving the connectivity;
12:    // Remove selected nodes
13:    Remove node  $\mathcal{X}$  from  $G'$ ;
14:   else
15:     // If no node is removed, copy a random node
16:     Copy a random node-edge pair  $(v_{copy}, e_{copy}) \in G'$ ;
17:     Connect a randomly selected node  $v_{orig}$  in  $G'$  to  $v_{copy}$  using  $e_{copy}$ ;
18:   end if
19:   // Store the modified graph
20:   Add  $G'$  to set  $\mathcal{G}$ ;
21: end for
22: Return  $\mathcal{G} = \{G'_1, G'_2, \dots, G'_s\}$ 

```

a post-training step following supervised learning in reasoning tasks like code generation (Hui et al., 2024; Le et al., 2022; Weyssow et al., 2024). These methods rely on high-quality labels to construct preference pairs (Pace et al., 2024). However, the available preference pairs are virtually nonexistent in domain-specific applications, particularly in underexplored areas such as VPLs. To mitigate this limitation, we propose a graph editing approach to construct preference pairs.

Negative sampling via graph editing As described in §2.2, we transform visually structured programs into NetworkX graphs to facilitate access to nodes and edges. Based on this graphical representation, we introduce a method for collecting preference pairs. We collect negative samples through graph editing, as described in Algorithm 1. Given a graph G corresponding to ground-truth

VPL code, we randomly delete nodes and edges, with a deletion ratio τ set to 0.1. This modifies approximately 10% of the graph structure. For graphs with fewer than 10 nodes, no deletions are performed. Instead, as shown in line 15, we augment the graph by duplicating and adding a node-edge pair. To ensure diversity, we generate 10 negative samples for each graph using different random seeds. The effect of varying τ is analyzed in §6.2.

Hard negative selection Preference optimization using hard negative samples is beneficial for learning subtle differences that supervised learning alone may not capture (Zhu et al., 2024; Chen et al., 2024). This suggests that the difference in connectivity between hub nodes and leaf nodes is substantial, and when nodes are randomly deleted, the quality of negative data varies considerably. Therefore, we select hard negative samples from the negative set collected via graph editing. To do so, we utilize the Graph Edit Distance (GED) (Sanfeliu and Fu, 1983; Abu-Aisheh et al., 2015), which quantifies the structural difference between graphs. Specifically, we use the GED-based metric for selection:

$$\text{GED}(G, G') = \min_{T \in \mathcal{T}(G, G')} \sum_{e \in T} \text{cost}(e) \quad (3)$$

where $\mathcal{T}(G, G')$ is the set of all possible edit operation that transform G' into G . Given $\mathcal{T}(G, G')$ and the cost, the GED is the minimum associative cost of each edit operation across all possible edit paths. In this study, the cost of an edit path is measured using the Levenshtein distance (Black, 1998). To construct preference pairs, we select the graph with the lowest edit distance to the reference graph from $\mathcal{G} = \{G'_1 \dots G'_s\}$ collected through multiple seeds. We designate the graph with the lowest score as the hard negative G_{Hard} . Finally, we construct a preference pair by pairing ground-truth graph G with G_{Hard} .

Direct preference optimization Preference learning is conducted using the selected preference pairs, with the retrieved prompt and its corresponding VPL code augmented as input. Implementation details are provided in Appendix B.

4 Experimental Settings

Dataset Due to the lack of publicly available datasets, we created our own by annotating ladder diagrams from actual production environments.

Using XG5000 (XG5, 2009), we exported these diagrams as XML files and divided them into functional units, where each unit consists of one or more interconnected rungs designed to perform a specific function. An experienced PLC programmer then annotated natural language instructions for each unit. The dataset was randomly split into training, validation, and test sets of 13,124, 500, and 500 samples. 80% of the training data was used for SFT, while the remaining 20% was used for preference learning. Appendix C shows detailed information about the dataset utilized in our study.

Implementation details We utilize Llama-3.1-8B-Instruct (Dubey et al., 2024) as the main backbone model and also use Qwen2.5-7B-Instruct (Yang et al., 2024) to assess the generalizability of our method. To facilitate task understanding, we provide a detailed task description and explanations of the visual elements used in ladder diagrams via the system prompt. For models using retrieval, we utilize BM25 (Robertson et al., 2009), which is a widely used lexical matching-based method. Using BM25, we augment the input to these retrieval-based models with a top-1 retrieved prompt-code pair from the training dataset³.

5 Main Results

Stage 1 (RAFT-V) improves upon SFT Table 2 compares the performance of SFT, RAFT-V, and our two-stage method across different output formats (XML, JSON, metaprogram) with two backbone models. By incorporating retrieval augmentation, RAFT-V consistently outperforms SFT and significantly enhances VPL generation. For example, in the JSON format, RAFT-V raises the Program EM from 52.6% to 63.2%, demonstrating more accurate VPL generation. Similar improvements are observed for other metrics.

Preference optimization (Stage 2) yields further gains Building on Stage 1’s improvements, Stage 2 further boosts correctness through preference optimization, particularly in terms of exact match (EM) scores. As shown in Table 2, our two-stage approach improves Program EM over RAFT-V by 3.4% in the metaprogram format and achieves an overall 13.2% gain compared to SFT. Because EM only assigns a score when the entire graph exactly matches, even minor generation failures can sig-

³System prompts are in Appendix G, and further implementation details are in Appendix B.

Format	Method	Node F1	Edge F1	Node EM	Edge EM	Program EM
XML	SFT	84.8 / 79.0	74.7 / 68.0	55.2 / 46.4	54.8 / 45.6	54.8 / 45.6
	RAFT-V	88.1 / 87.4	80.9 / 78.7	63.8 / 59.6	62.8 / 58.8	62.8 / 58.8
	Ours	89.6 / 87.4	82.6 / 79.0	66.2 / 61.2	65.2 / 60.0	65.2 / 60.0
JSON	SFT	85.1 / 81.2	74.9 / 69.3	53.6 / 46.0	52.6 / 45.4	52.6 / 45.4
	RAFT-V	90.1 / 88.4	82.0 / 79.6	63.8 / 59.2	63.2 / 57.4	63.2 / 57.4
	Ours	90.7 / 88.7	83.0 / 80.5	66.0 / 61.6	65.2 / 60.2	65.2 / 60.2
Metaprogram	SFT	86.3 / 81.7	75.7 / 70.2	55.2 / 47.6	54.0 / 47.2	54.0 / 47.2
	RAFT-V	89.9 / 89.1	82.4 / 80.4	64.6 / 60.6	63.8 / 59.6	63.8 / 59.6
	Ours	90.6 / 89.7	83.6 / 81.6	68.6 / 63.2	67.2 / 61.8	67.2 / 61.8

Table 2: Main results of the 2-stage training strategy. Each score is presented in the order of Llama-3.1-8B-Instruct / Qwen2.5-7B-Instruct.

k	Node F1	Edge F1	Node EM	Edge EM	Program EM
1 (Ours)	89.9	82.4	64.6	63.8	63.8
2	90.6	83.5	66.4	65.6	65.6
3	90.9	84.0	65.4	64.6	64.6

Table 3: Performance of RAFT-V across different numbers of retrieved examples k . We use the Llama3.1-8B-Instruct model with metaprogram format.

nificantly impact the metric. These results show that preference optimization reduces such minor generation errors and ensures more precise outputs. Notably, these gains are consistent across all output formats and base models, demonstrating the robustness of our approach.⁴ Since the generated PLC code is directly deployed in factory environments, careful review by programmers is still essential. In this context, a more than 10% increase in Program EM scores across all output formats compared to SFT is practically significant in real industrial settings. This translates to faster validation cycles and reduced manual effort, making the overall development process more efficient and cost-effective.

6 Discussion

In this section, we evaluate RAFT-V under different retrieved examples (§6.1), analyze the deletion ratio in graph editing (§6.2), compare graph sampling based on graph editing to an existing sampling strategy (§6.3), and investigate the impact of program complexity on our approach (§6.4).

6.1 Impact of Number of Retrieved Examples

To assess whether the number of retrieved examples affects performance, we evaluate RAFT-V trained with different numbers of retrieval examples $k \in \{1, 2, 3\}$ (Table 3). Although increasing

⁴See Appendix D for human evaluation details.

retrieval examples k generally improves generation quality, the gains are marginal, which indicates that even a few examples are able to capture functional patterns in visual programming languages. Based on this observation, we set $k = 1$ for our main experiments.

6.2 Variation in Deletion Ratio

τ	Node F1	Edge F1	Node EM	Edge EM	Program EM
0	90.7	83.4	67.2	65.8	65.8
0.1 (Ours)	90.6	83.6	68.6	67.2	67.2
0.5	90.5	83.3	68.2	66.8	66.8
0.9	90.6	83.4	67.4	66.0	66.0

Table 4: Impact of varying deletion ratios. Results for Llama-3.1-8B-Instruct model using metaprogram format.

Furthermore, we report the performance variation depending on the deletion ratio τ . We train the model using preference pairs with different τ values, where $\tau \in \{0, 0.1, 0.5, 0.9\}$. Table 4 shows that F1 scores remain stable across τ values, while EM scores decrease as τ increases. In particular, when τ is 0.9, the negative samples during preference training are graphs with 90% of the original graph removed. As a result, the model can easily distinguish them, making the training less effective. Based on these results, we set τ to 0.1.

Methods	Node F1	Edge F1	Node EM	Edge EM	Program EM
RAFT-V (100%)	90.6	83.6	66.6	65.2	65.2
RAFT-V	89.9	82.4	64.6	63.8	63.8
BoN (Random)	90.4	83.2	66.8	65.6	65.6
BoN (Unstrict)	90.7	83.5	67.0	65.8	65.8
BoN (Strict)	90.5	83.3	67.4	66.2	66.2
Ours (Random)	90.6	83.5	68.0	66.6	66.6
Ours	90.6	83.6	68.6	67.2	67.2

Table 5: Comparison of BoN sampling and graph editing. Results for metaprogram using Llama-3.1-8B-Instruct.

6.3 Analysis of Graph Editing

We validate our graph editing approach for collecting preference pairs by comparing it with Best-of-N (BoN) sampling⁵ (Stiennon et al., 2020; Snell et al., 2025), which selects the best generation among N candidates. We generate $N = 10$ outputs from the RAFT-V model and compare them with our negative candidate set $\mathcal{G} = \{G'_1 \dots G'_{10}\}$, which is derived from graph editing (algorithm 1).

Preference learning is effective To address the concern that the gains in the 2-stage approach may result from the additional data used for preference learning, we compare it with a 1-stage approach, where the model is trained on all the data without a second stage and is referred to as RAFT-V (100%). As shown in Table 5, RAFT-V improves performance as the dataset size increases. However, RAFT-V (100%) shows only marginal improvement in EM scores compared to preference-learned models. Although RAFT-V (100%) demonstrates effectiveness in improving F1 scores when compared to BoN-based preference learning baselines, it is insufficient to minimize minor errors (EM).

Editing-based negative selection is efficient We introduce four baselines for comparison with graph editing: BoN (Random), where \mathcal{G} is sampled from BoN, and G_{Hard} is randomly selected from BoN-sampled \mathcal{G} ; BoN (Unstrict), which selects preference pairs based on GED; and BoN (Strict), which considers preference pairs for training only when the sampled output includes an exact match with the correct answer (i.e., GED = 0). Finally, in the case of Ours (Random), \mathcal{G} is sampled from graph editing, but G_{Hard} is randomly selected instead of using GED.

As shown in Table 5, BoN (Strict) constructs preference pairs based on exact matches and achieves the highest accuracy among BoN-based methods despite utilizing only 30% of the data. This result demonstrates that dataset quality has a more significant impact on preference learning than dataset size (Hou et al., 2024; Kim et al., 2024). However, BoN sampling often fails to generate challenging cases consistently, as negative samples are selected from the sampled outputs. In contrast, our editing-based negative selection provides a more systematic approach to generating negative

⁵For BoN sampling, we use nucleus decoding (Holtzman et al., 2020) with temperature=1.0, top_p=1.0.

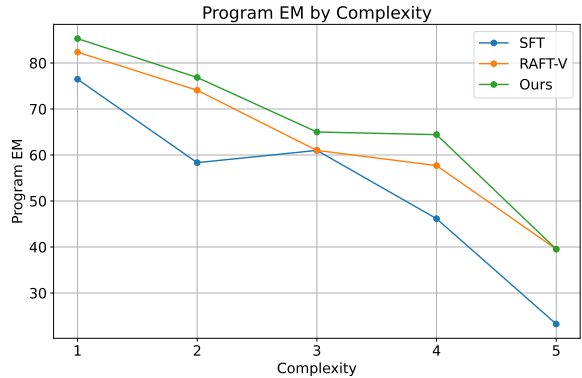


Figure 4: Program EM score across different complexities. We use the metaprogram format with the Llama3.1-8B-Instruct model.

samples, which consistently enables the generation of hard negative pairs with higher efficiency.

6.4 Performance Across Program Complexity

To evaluate our methodology’s performance across varying program complexities, we convert each program (VPL code) in the test split into NetworkX graphs. We define complexity as the total number of nodes and edges in the graph. We then sort programs by complexity and split them into five percentile ranges: 0–20%, 20–40%, 40–60%, 60–80%, and 80–100%, labeling them from 1 to 5. Figure 4 shows the average Program EM across these categories. Our approach consistently outperforms the SFT baseline, widening the gap at higher complexities (+18.3% in 4, +16.3% in 5). These findings highlight our method’s robustness, with benefits increasing in challenging scenarios.

7 Related Work

LLMs for Visual Program Generation Visual programming systems (e.g., LabView (Bitter et al., 2006), XG5000 (XG5, 2009)) typically feature node-based interfaces that let users visually write and modify programs. Recently, researchers have begun utilizing LLMs to generate VPLs, as they are known for their powerful text-based code generation capabilities. For example, Cai et al. (2024) integrates low-code visual programming with LLM-based task execution for direct interaction with LLMs, while Zhang et al. (2024b) studies generation of node-based visual dataflow languages in audio programming. Similarly, Xue et al. (2024); Zhou et al. (2025) investigates Machine Learning workflow generation from natural language commands and demonstrates that metaprogram-based

text formats outperform other formats like JSON. However, these prompting-based methods face limitations for VPLs like Ladder Diagram, where custom I/O mapping and domain-specific syntax are crucial. Thus, we study fine-tuning approaches with domain-specific data to better capture these details.

LLM-based PLC code generation Programmable Logic Controllers (PLCs) are essential components in industrial automation and are used to control machinery and processes reliably and efficiently. Among the programming languages defined by the IEC 61131-3 standard (IEC, 2013), Structured Text (ST) and Ladder Diagram (LD) are commonly used for programming PLCs. Research in this area has focused on utilizing LLMs to generate ST code from natural language descriptions. Recent studies have demonstrated the potential of LLMs in generating high-quality ST code (Koziolok et al., 2023, 2024), enhancing safety and accuracy with verification tools and user feedback (Fakih et al., 2024), and automating code generation and verification using multi-agent frameworks (Liu et al., 2024). Although these advances have improved PLC code generation, they primarily focus on ST, despite LD being widely used in industrial settings due to its similarity to electrical circuits (Malik, 2024). While Zhang and de Sousa (2024) attempts to generate LD as an ASCII art based on user instructions in a zero-shot manner, their findings show that even advanced LLMs struggle with basic LD generation. These limitations highlight the necessity of training-based methods for LD generation. In this work, we address this gap by introducing a training-based approach for LLMs to generate LD and thus pave the way for the broader adoption of AI-assisted PLC programming.

8 Conclusion

In this paper, we have demonstrated the importance of fine-tuning for VPL generation and introduced a novel two-stage training strategy. By combining retrieval augmentation leveraging repetitive sub-routines in VPLs with preference learning through graph editing, we achieved significant improvements in LD generation. Our work marks a crucial step toward LLM-based LD generation, and given that the method leverages VPL characteristics, it holds potential for broader applicability.

Limitations

One limitation of our study is that the experiments were conducted exclusively on ladder diagrams. Although ladder diagrams are widely used in visual programming languages (VPLs), extending our methodology to other VPLs is necessary for broader applicability. Nonetheless, ladder diagrams remain crucial in industrial automation but have been largely overlooked due to the text-centric design of most large language models (LLMs). Prior studies have struggled to generate even basic ladder diagrams, highlighting a significant gap. To bridge this, we proposed a method to represent ladder diagrams in multiple text formats, enabling LLMs to process them as a graphical language. This approach addresses the limitations of previous studies and opens new possibilities for LLM-based generation of visual programming languages.

Another limitation is that our dataset cannot be publicly released due to strict confidentiality constraints, as it contains proprietary industrial processes used in manufacturing. To safeguard sensitive information, we have anonymized and provided only a partial subset. Our dataset adhered to the IEC 61131-3 international standard and specifically utilized ladder diagrams, which are a widely adopted language in PLC programming. While our dataset has unique attributes for ladder elements and rung structures that differ across users, it remains applicable to other ladder diagram-based programs that adhere to the IEC 61131-3 standard. Consequently, our methodology is not restricted to this particular dataset but rather is applicable to other industrial formats, which demonstrates scalability and adaptability. Yet, recognizing the importance of further generalization, we acknowledge the need for a publicly available VPL-based benchmark, which we leave as future work.

Ethical Considerations

In our research, we employed models such as Llama-3.1-8B-Instruct and Meta-Llama-3.1-70B-Instruct-AWQ-INT4, both of which are released under the Llama 3.1 Community License, as well as Qwen2.5-7B-Instruct, released under the Apache License 2.0, and Qwen2.5-72B-Instruct-AWQ, released under the Qwen license. All models were used strictly for research purposes, and no artifacts were utilized beyond the scope of the study.

Acknowledgments

This work was supported by Hyundai Mobis (47.5%). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2025-2020-0-01789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation, 47.5%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH), 5%).

References

2009. *XG5000 User Manual*. LS Electric. <https://cdn.automationdirect.com/static/manuals/lselectric/xg5000usermanual.pdf>.
2013. IEC 61131-3: Programmable controllers – Part 3: Programming languages.
- Zeina Abu-Aisheh, Romain Raveaux, Jean-Yves Ramel, and Patrick Martineau. 2015. An exact graph edit distance algorithm for solving pattern recognition problems. In *4th International Conference on Pattern Recognition Applications and Methods 2015*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ephrem Ryan Alphonsus and Mohammad Omar Abdullah. 2016. A review on the applications of programmable logic controllers (plcs). *Renewable and Sustainable Energy Reviews*, 60:1185–1205.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. *Program synthesis with large language models*. *Preprint*, arXiv:2108.07732.
- Rick Bitter, Taqi Mohiuddin, and Matt Nawrocki. 2006. *LabVIEW: Advanced programming techniques*. Crc Press.
- Paul E Black. 1998. Dictionary of algorithms and data structures.
- Yuzhe Cai, Shaoguang Mao, Wenshan Wu, Zehua Wang, Yaobo Liang, Tao Ge, Chenfei Wu, WangYou WangYou, Ting Song, Yan Xia, Nan Duan, and Furu Wei. 2024. *Low-code LLM: Graphical user interface over large language models*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 12–25, Mexico City, Mexico. Association for Computational Linguistics.
- Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2023. *Codet: Code generation with generated tests*. In *The Eleventh International Conference on Learning Representations*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. *On softmax direct preference optimization for recommendation*. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Christian DeLozier and James Shey. 2023. Using visual programming games to study novice programmers. *International Journal of Serious Games*, 10(2):115–136.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Kelvin T Erickson. 1996. Programmable logic controllers. *IEEE potentials*, 15(1):14–17.
- Mohamad Fakh, Rahul Dharmaji, Yasamin Moghaddas, Gustavo Quiros, Oluwatosin Ogundare, and Mohammad Abdullah Al Faruque. 2024. Llm4plc: Harnessing large language models for verifiable programming of plcs in industrial control systems. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, pages 192–203.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Epic Games. 2010. Blueprints: Visual scripting in unreal engine. <https://dev.epicgames.com/documentation/en-us/unreal-engine/blueprints-visual-scripting-in-unreal-engine>. Accessed: 2025-02-10.
- Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. *Deepseek-coder: When the large language model meets programming – the rise of code intelligence*. *Preprint*, arXiv:2401.14196.

- Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text de-generation](#). In *International Conference on Learning Representations*.
- Zhenyu Hou, Pengfan Du, Yilin Niu, Zhengxiao Du, Aohan Zeng, Xiao Liu, Minlie Huang, Hongning Wang, Jie Tang, and Yuxiao Dong. 2024. [Does rlhf scale? exploring the impacts from data, model, and method](#). *Preprint*, arXiv:2412.06000.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. [Qwen2.5-coder technical report](#). *Preprint*, arXiv:2409.12186.
- Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2024. [Aligning large language models with self-generated preference data](#). *Preprint*, arXiv:2406.04412.
- Heiko Kozirolek, Sten Gruener, and Virendra Ashiwal. 2023. Chatgpt for plc/dcs control logic generation. In *2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–8. IEEE.
- Heiko Kozirolek, Sten Grüner, Rhaban Hark, Virendra Ashiwal, Sofia Linsbauer, and Nafise Eskandani. 2024. Llm-based and retrieval-augmented control code generation. In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 22–29.
- Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven C.H. Hoi. 2022. Coderl: mastering code generation through pretrained models and deep reinforcement learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliachko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muftasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. [StarCoder: may the source be with you!](#) *Preprint*, arXiv:2305.06161.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Weiming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. [Awq: Activation-aware weight quantization for on-device llm compression and acceleration](#). In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.
- Zihan Liu, Ruinan Zeng, Dongxia Wang, Gengyun Peng, Jingyi Wang, Qiang Liu, Peiyu Liu, and Wenhao Wang. 2024. Agents4plc: Automating closed-loop plc code generation and verification in industrial control systems using llm-based agents. *arXiv preprint arXiv:2410.14209*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Saliha Malik. 2024. [Ladder logic programming](#). Accessed: 2025-01-22.
- Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. [West-of-n: Synthetic preferences for self-improving reward models](#). *Preprint*, arXiv:2401.12086.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Partha Pratim Ray. 2017. A survey on visual programming languages in internet of things. *Scientific Programming*, 2017(1):1231430.
- Pengzhen Ren, Min Li, Zhen Luo, Xinshuai Song, Ziwei Chen, Weijia Liufu, Yixuan Yang, Hao Zheng, Rongtao Xu, Zitong Huang, Tongsheng Ding, Luyang Xie, Kaidong Zhang, Changfei Fu, Yang Liu, Liang Lin, Feng Zheng, and Xiaodan Liang. 2024. [Infinite-world: A unified scalable simulation framework for general visual-language robot interaction](#). *Preprint*, arXiv:2412.05789.
- Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

- Alberto Sanfeliu and King-Sun Fu. 1983. [A distance measure between attributed relational graphs for pattern recognition](#). *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):353–362.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. [Scaling test-time compute optimally can be more effective than scaling LLM parameters](#). In *The Thirteenth International Conference on Learning Representations*.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Miguel Terra-Neves, João Nadkarni, Miguel Ventura, Pedro Resende, Hugo Veiga, and António Alegria. 2021. [Duplicated code pattern mining in visual programming languages](#). In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE '21*. ACM.
- Martin Weysow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. 2024. [Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences](#). *Preprint*, arXiv:2403.09032.
- Xiangyuan Xue, Zeyu Lu, Di Huang, Zidong Wang, Wanli Ouyang, and Lei Bai. 2024. [Comfybench: Benchmarking llm-based agents in comfyui for autonomously designing collaborative ai systems](#). *Preprint*, arXiv:2409.01392.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024a. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.
- William Zhang, Maria Leon, Ryan Xu, Adrian Cardenas, Amelia Wissink, Hanna Martin, Maya Srikanth, Kaya Dorogi, Christian Valadez, Pedro Perez, et al. 2024b. Benchmarking llm code generation for audio programming with visual dataflow languages. *arXiv preprint arXiv:2409.00856*.
- Yimin Zhang and Mario de Sousa. 2024. [Exploring llm support for generating iec 61131-3 graphic language programs](#). In *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*, page 1–7. IEEE.
- Zhongyi Zhou, Jing Jin, Vrushank Phadnis, Xiuxiu Yuan, Jun Jiang, Xun Qian, Jingtao Zhou, Yiyi Huang, Zheng Xu, Yinda Zhang, Kristen Wright, Jason Mayes, Mark Sherwood, Johnny Lee, Alex Olwal, David Kim, Ram Iyengar, Na Li, and Ruofei Du. 2025. Instructpipe: Building visual programming pipelines in visual blocks with human instructions using llms. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI)*, page 23.
- Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024. [Self-supervised visual preference alignment](#). In *ACM Multimedia 2024*.

A An Example of Metric Evaluations

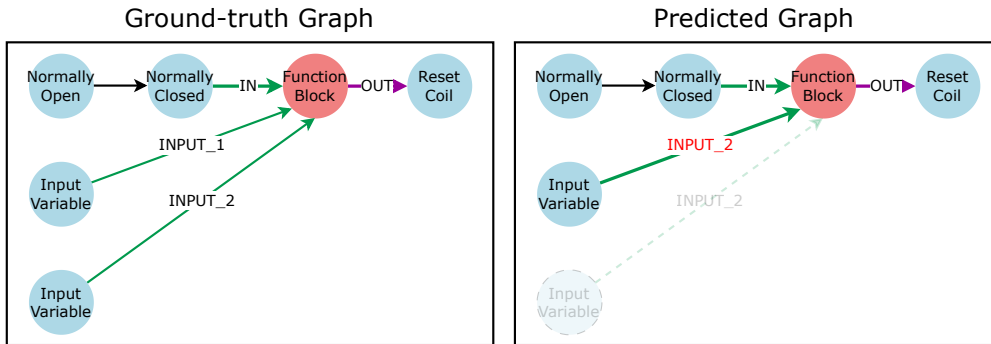


Figure 5: Example illustrating the evaluation metrics. Node and edge attributes have been modified from the original data due to security concerns.

Figure 5 presents the ground-truth and predicted graphs used for the following evaluation example. Based on these graphs, we illustrate how the proposed metrics capture both partial and exact correctness of the predicted Ladder Diagram (LD) program. First, consider the **node-level evaluation**. The ground-truth graph $G = (V, E)$ contains 6 nodes, while the predicted graph $\hat{G} = (\hat{V}, \hat{E})$ contains 5 nodes. All 5 predicted nodes correctly match with nodes in V , including all required attributes such as type and name. Hence,

$$\text{Precision}_N = \frac{|V \cap \hat{V}|}{|\hat{V}|} = \frac{5}{5} = 1.0, \quad \text{Recall}_N = \frac{|V \cap \hat{V}|}{|V|} = \frac{5}{6} = 0.833$$

Using these values, the Node F1 score is:

$$\text{Node F1} = \frac{2 \times 1.0 \times 0.833}{1.0 + 0.833} = 0.91$$

Next, we examine the **edge-level evaluation**. The reference graph has 5 edges, while the predicted graph contains only 4. Among the predicted edges, 3 match the ground-truth edges exactly in terms of connected nodes and attributes. Therefore,

$$\text{Precision}_E = \frac{3}{4} = 0.75, \quad \text{Recall}_E = \frac{3}{5} = 0.6$$

$$\text{Edge F1} = \frac{2 \times 0.75 \times 0.6}{0.75 + 0.6} = 0.67$$

Finally, we assess **Exact Match (EM)**. Since both the node and edge sets do not exactly match those in the reference graph—one node is missing and one edge is incorrect—all EM scores are 0:

$$\text{Node EM} = 0, \quad \text{Edge EM} = 0, \quad \text{Program EM} = 0$$

B Further Implementation Details

Supervised fine-tuning (including RAFT-V) We fine-tuned models using SFT with 10 epochs, a batch size of 8, and a learning rate of 5×10^{-5} . We applied LoRA (Hu et al., 2022) with a rank of 256, $\alpha = 256$, and a dropout rate of 0.05. LoRA adaptation was applied to the following target modules: q_proj, v_proj, k_proj, o_proj. Using the AdamW (Loshchilov and Hutter, 2019) optimizer, we minimized the cross-entropy loss \mathcal{L}_{CE} between ground-truth code c and the predicted code \hat{c} as follows:

$$\mathcal{L}_{\text{CE}} = - \sum_{t=1}^T \sum_{v \in V} c_t(v) \log \hat{c}_t(v)$$

where T is the sequence length and V is a vocabulary size. The SFT process took 6 hours using 4 A100-80GB GPUs, while the SFT process for RAFT-V took 8 hours using the same hardware configuration.

Preference learning For preference learning, we utilized Direct Preference Optimization (DPO) (Rafailov et al., 2023) and trained models for 5 epochs with a batch size of 64. The learning rate was set to 1×10^{-7} , with a warmup ratio of 0.03, a weight decay of 0.01, and $\beta = 0.1$. As with SFT, LoRA was applied with the same rank, α , dropout rate, and target modules.

Given a pair of responses, a preferred response y^+ and a dispreferred response y^- for a given input x , we minimize the following DPO loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y^+, y^-) \sim \mathcal{D}} \left[\log \sigma \left(\beta \left(\log \frac{\pi_{\theta}(y^+ | x, \mathcal{R}(x))}{\pi_{\text{ref}}(y^+ | x, \mathcal{R}(x))} - \log \frac{\pi_{\theta}(y^- | x, \mathcal{R}(x))}{\pi_{\text{ref}}(y^- | x, \mathcal{R}(x))} \right) \right) \right]$$

where $\sigma(z)$ is the sigmoid function, β is a scaling factor controlling the strength of preference optimization, $\mathcal{R}(x) = (p^r, c^r)$ represents a retrieved prompt and its corresponding VPL code obtained from a retriever \mathcal{R} , and \mathcal{D} represents the dataset of preference-labeled samples. The reference model π_{ref} serves as a baseline to prevent reward overoptimization, ensuring stable preference learning. The preference training stage took 6 hours on 4 A100-80GB GPUs. We measured the loss on the validation set at each epoch in both training stages and applied early stopping based on this criterion, with a patience value of 2.

Sampling parameters We employed beam search decoding with a beam size of 4 during inference.

C Dataset Details

Type	Content
Prompt	<p>Program Description: 셀 이재기 서보 조그 고속 하한 값 인터락 프로그램을 만들어줘. (Create a cell transfer servo jog high-speed interlock program.)</p> <p>Detailed Description: 조그 고속 속도 설정값이 20 이하일 때 조그 고속속도를 20으로 설정해줘. (When the jog high-speed setting value is 20 or less, set the jog high-speed to 20.)</p>
XML (Anonymized)	<pre><Rung> <Element Attributes..., Coordinate="X"></Element> <Element Attributes..., Coordinate="Y"></Element> <Element Attributes..., Coordinate="Z"></Element> </Rung></pre>

Table 6: Program description and anonymized XML example.

Process Type	Count	Ratio (%)
Cell Distribution	2,893	20.5
Cell Processing	3,093	21.9
Cell Vision Inspection	726	5.1
Cell Stacking	3,706	26.2
Pack End-plate Supply	2,808	19.9
Other Processes	898	6.4

Table 7: Distribution of process types in the dataset.

An example from the dataset used in this study is provided. The dataset was extracted from ladder logic (LD) programs deployed in real-world battery cell manufacturing facilities, and consists of **prompt** and **code** pairs. The prompt was collected in Korean and used without translation for model training and evaluation. The prompt is further divided into **Program Description**, providing a high-level summary of the functionality, and **Detailed Description**, specifying task parameters and conditions. The prompts were created by an experienced PLC programmer from the Republic of Korea, ensuring the incorporation of

domain expertise into the collected data. The programmer was explicitly informed that the dataset would be collected for model training and evaluation and used strictly for research purposes. We obtained consent prior to data collection. The `code` is represented in XML format, which describes the ladder diagram with elements. These elements are listed sequentially, and each element includes its attributes (e.g., variable names, parameters) and coordinate information. The datasets were constructed by an experienced PLC engineer, and all sensitive information was anonymized to ensure data confidentiality.

In addition, we report the distribution of process types included in the dataset. We constructed fine-grained and diverse operational data across multiple stages of production, including cell distribution, processing, stacking, vision inspection, and pack component supply. The statistics of the dataset for each stage are presented in Table 7. We hope that this statistical information will serve as a valuable reference for future research.

D Human Evaluation

To further evaluate the effectiveness of our methods, we conduct human evaluations on a randomly selected 50 test set examples. Due to the dataset’s security sensitivity, human evaluations were conducted exclusively by PLC engineers who had authorized access to confidential information. Three experienced LD programmers evaluated the generated code based on functionality and assigned scores on a scale of 1 to 5. The evaluation process is as follows: First, we convert the generated ladder program in each text format into an actual Ladder Diagram that can be opened in the IDE (XG5000). Then, we manually capture both the reference and predicted ladder programs from the IDE. Finally, the evaluators are asked to assess each reference–prediction screenshot pair according to the criteria. The criteria for scoring are reported in Table 8. The ratings were based on their professional experience and the practical applicability of the code in real industrial settings.

The results of the human evaluation are shown in Table 9. Compared to SFT-only, RAFT-V showed an improvement of 0.26 points, while our method outperformed RAFT-V by 0.02 points. Furthermore, we report Fleiss’ kappa coefficient (Fleiss, 1971) to statistically evaluate the level of agreement among human evaluators. The results indicate that our proposed methodology demonstrates the highest degree of inter-rater consistency.

Score	Assessment Criteria
5	Immediately usable in real-world deployment.
4	Mostly appropriate; minor polishing may help.
3	Valid but awkward; manual revision is needed.
2	Mostly correct syntax but clearly wrong logic or intent.
1	Unexecutable or severely broken structure with partial syntax validity.

Table 8: Scoring guidelines for human evaluation.

Method	Functional Score \uparrow	Kappa Score \uparrow
SFT-only	4.34	0.62
RAFT-V	4.60	0.58
Ours	4.62	0.66

Table 9: Average of human evaluation results. We used the metaprogram format for evaluation.

E More Details on RAG

The prompt template for RAG models is as follows. The RAG model takes the same system prompt as the training-based model depending on the text format.

```

message = [
  {"role": "system", "content": {system_prompt}},
  {"role": "user", "content": retrieved_prompt},
  ...
  {"role": "assistant", "content": retrieved_code},
  {"role": "user", "content": {test_prompt}}
]

```

Figure 6: Prompt template for RAG models.

Since some outputs of RAG models were ill-formed (particularly in the case of Qwen), we applied postprocessing to refine them. Extra elements such as ``xml``, ``json``, ``python``, or unrelated code snippets were removed. For inference, we employ beam search with beam size of 4.

F Detailed Results

In this section, we present detailed experimental results. Table 10 shows the performance comparison between the SFT model and RAG, with a larger LLM within the same family.

Format	Method	Node F1	Edge F1	Node EM	Edge EM	Program EM
XML	SFT	84.8 / 79.1	74.7 / 68.2	55.2 / 46.8	54.8 / 46.0	54.8 / 46.0
	$N = 0$	1.2 / 1.0	0.5 / 0.0	0.0 / 0.0	0.4 / 0.0	0.0 / 0.0
	$N = 1$	69.0 / 19.5	58.2 / 13.9	39.0 / 8.8	38.6 / 8.6	38.0 / 8.6
	$N = 3$	76.3 / 61.4	67.6 / 53.3	49.8 / 40.2	49.0 / 39.4	49.0 / 39.4
	$N = 5$	74.4 / 60.7	67.1 / 53.8	49.6 / 41.2	49.2 / 40.6	49.2 / 40.6
	$N = 7$	59.3 / 53.6	53.5 / 47.9	42.8 / 37.8	42.2 / 37.6	42.2 / 37.6
	$N = 9$	50.5 / 44.9	45.1 / 40.4	37.6 / 34.4	36.8 / 34.0	36.8 / 34.0
JSON	SFT	85.1 / 81.2	74.9 / 69.3	53.6 / 46.8	52.6 / 46.4	52.6 / 46.4
	$N = 0$	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
	$N = 1$	1.2 / 5.3	0.8 / 3.9	0.6 / 3.6	0.6 / 4.0	1.1 / 3.6
	$N = 3$	79.2 / 73.4	69.8 / 62.7	48.2 / 43.4	47.4 / 46.0	47.4 / 46.0
	$N = 5$	79.3 / 74.8	70.7 / 67.1	51.4 / 43.8	50.6 / 46.6	50.6 / 46.6
	$N = 7$	72.6 / 54.6	65.7 / 50.2	51.6 / 42.2	50.8 / 41.8	50.8 / 41.8
	$N = 9$	58.5 / 52.8	53.1 / 48.0	44.2 / 40.2	43.6 / 39.8	43.6 / 39.8
Metaprogram	SFT	86.3 / 81.8	75.7 / 70.5	55.2 / 48.4	54.0 / 48.0	54.0 / 48.0
	$N = 0$	0.7 / 0.0	0.0 / 0.0	0.0 / 0.0	0.2 / 0.0	0.0 / 0.0
	$N = 1$	4.2 / 1.0	3.5 / 0.9	2.8 / 0.8	2.8 / 0.6	2.8 / 0.6
	$N = 3$	80.8 / 47.2	71.1 / 40.9	49.4 / 27.0	48.4 / 26.0	48.4 / 26.0
	$N = 5$	80.0 / 68.1	71.9 / 59.7	52.2 / 43.6	51.8 / 42.6	51.8 / 42.6
	$N = 7$	76.4 / 71.5	69.1 / 64.2	51.8 / 48.2	51.2 / 47.6	51.2 / 47.6
	$N = 9$	64.5 / 59.5	57.8 / 53.6	47.2 / 43.8	46.8 / 42.8	46.8 / 42.8

Table 10: Performance comparison between SFT and RAG (with a larger LLM) across XML, JSON, and Metaprogram formats. For RAG, N denotes the number of retrieved examples. Each SFT score is presented in the order of "Llama-3.1-8B-Instruct / Qwen2.5-7B-Instruct." RAG model utilizes the AWQ-quantized versions of [hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4](#) and [Qwen/Qwen2.5-72B-Instruct-AWQ](#). The highest performance value within each format is shown in bold.

G System Prompts

Depending on the type of text format used, the model takes a different system prompt. The retrieval-based model used in the experiments takes the following prompt template as input:

```

message = [
  {"role": "system", "content": {system_prompt}},
  {"role": "user", "content": retrieved_prompt},
  ...
]

```



```

{"role": "assistant", "content": retrieved_code},
{"role": "user", "content": f"Based on the given input, generate the corresponding code:
{test_prompt}"}
]

```

Figure 7: Prompt template used in this study.

For models that do not utilize retrieval, the prompt template excludes `retrieved_prompt` and `retrieved_code` for code generation.

G.1 System prompt for XML

```

You are a programming assistant specializing in generating ladder programs in XML format.
Your task is to translate functional descriptions into equivalent PLC ladder logic and
directly represent the ladder logic as XML. The natural language instructions will
describe the desired functionality. Your job is to:
1. Interpret the described functionality.
2. Translate it into equivalent ladder logic components (e.g., rungs, contacts, coils).
3. Directly create and output the ladder logic as XML.

### Requirements for Ladder Logic Representation in XML:
- Each element must include an `ElementType` attribute, which specifies its type, and
  additional necessary attributes depending on the `ElementType`:
- The output XML must be well-formed, human-readable, and valid for parsing by PLC-related
  tools or frameworks.

### Explanation of ElementTypes:
[Lines]
- VertLine: It is a vertical line.
- HorzLine: It is a horizontal line.
- MultiHorzLine: It is a horizontal line with a fixed length.

[Contact]
- NormallyOpen: When the state of the BOOL variable (indicated by "***") is On, the state of
  the left connection line is copied to the right connection line. Otherwise, the state of
  the right connection line is Off.
- NormallyClosed: When the state of the BOOL variable (indicated by "***") is Off, the state
  of the left connection line is copied to the right connection line. Otherwise, the state
  of the right connection line is Off.
- RisingEdgeContact: If the value of the BOOL variable (indicated by "***") changes from Off
  in the previous scan to On in the current scan, and the state of the left connection line
  is On, the state of the right connection line becomes On during the current scan.
- FallingEdgeContact: If the value of the BOOL variable (indicated by "***") changes from On
  in the previous scan to Off in the current scan, and the state of the left connection
  line is On, the state of the right connection line becomes On during the current scan.
- RisingEdgeNotContact: If the value of the BOOL variable (indicated by "***") changes from
  Off in the previous scan to On in the current scan, and the state of the left connection
  line is On, the state of the right connection line becomes Off during the current scan.
- FallingEdgeNotContact: If the value of the BOOL variable (indicated by "***") changes from
  On in the previous scan to Off in the current scan, and the state of the left connection
  line is On, the state of the right connection line becomes Off during the current scan.

[Coil]
- StandardCoil: The state of the left connection line is assigned to the corresponding BOOL
  variable (indicated by "***").
- NegatedCoil: The negated value of the left connection line state is assigned to the
  corresponding BOOL variable (indicated by "***"). If the left connection line state is
  Off, the corresponding variable is set to On, and if the left connection line state is
  On, the corresponding variable is set to Off.
- SetCoil: When the state of the left connection line becomes On, the corresponding BOOL
  variable (indicated by "***") is set to On and remains On until turned Off by the Reset
  coil.
- ResetCoil: When the state of the left connection line becomes On, the corresponding BOOL
  variable (indicated by "***") is set to Off and remains Off until turned On by the Set
  coil.

```

- RisingEdgeCoil: If the state of the left connection line changes from Off in the previous scan to On in the current scan, the value of the corresponding BOOL variable (indicated by "***") becomes On only during the current scan.
 - FallingEdgeCoil: If the state of the left connection line changes from On in the previous scan to Off in the current scan, the value of the corresponding BOOL variable (indicated by "***") becomes On only during the current scan.
- [Others]
- Inverter: The state of the left connection line is inverted and passed to the right connection line.
 - FunctionBlock: Represents a function block.
 - Variable: Represents the variable corresponding to the function.
 - RisingEdge: Before detecting a positive transition, if the result of the previous operations changes from Off in the previous scan to On in the current scan, and the state of the left connection line is On, the state of the right connection line becomes On only during the current scan.
 - FallingEdge: Before detecting a negative transition, if the result of the previous operations changes from On in the previous scan to Off in the current scan, and the state of the left connection line is On, the state of the right connection line becomes On only during the current scan.

Figure 8: System prompt for XML.

G.2 System prompt for JSON

- You are a programming assistant specializing in generating ladder programs in JSON format. Your task is to translate functional descriptions into equivalent PLC ladder logic and directly represent the ladder logic as JSON. The natural language instructions will describe the desired functionality. Your job is to:
1. Interpret the described functionality.
 2. Translate it into equivalent ladder logic components (e.g., contacts, coils, functions).
 3. Directly create and output the ladder logic as JSON.
- ### Requirements for Ladder Logic Representation in JSON:
- The JSON structure must adhere to the following format:
 - The root is an object containing a single graph, such as ``G0``, which represents the ladder logic network.
 - Each node in the graph is identified by a unique ID (e.g., ``0``, ``9``, etc.).
 - Each node has:
 - ``attributes``: An object containing the properties of the node, including:
 - ``ElementType``: The type of ladder logic element (e.g., `"NormallyOpen"`, `"StandardCoil"`, `"Variable"`, `"FunctionBlock"`).
 - Additional attributes specific to the ``ElementType``
 - ``edges``: An array of connections from this node to other nodes, where:
 - Each edge has a ``target`` (the ID of the target node) and a ``type`` (the connection type, e.g., `"Enable"`, `"Output"`, `"Input1"`).
- [Contact]
- NormallyOpen: When the state of the BOOL variable (indicated by "***") is On, the state of the left connection line is copied to the right connection line. Otherwise, the state of the right connection line is Off.
 - NormallyClosed: When the state of the BOOL variable (indicated by "***") is Off, the state of the left connection line is copied to the right connection line. Otherwise, the state of the right connection line is Off.
 - RisingEdgeContact: If the value of the BOOL variable (indicated by "***") changes from Off in the previous scan to On in the current scan, and the state of the left connection line is On, the state of the right connection line becomes On during the current scan.
 - FallingEdgeContact: If the value of the BOOL variable (indicated by "***") changes from On in the previous scan to Off in the current scan, and the state of the left connection line is On, the state of the right connection line becomes On during the current scan.
 - RisingEdgeNotContact: If the value of the BOOL variable (indicated by "***") changes from Off in the previous scan to On in the current scan, and the state of the left connection line is On, the state of the right connection line becomes Off during the current scan.
 - FallingEdgeNotContact: If the value of the BOOL variable (indicated by "***") changes from On in the previous scan to Off in the current scan, and the state of the left connection line is On, the state of the right connection line becomes Off during the current scan.

```

[Coil]
- StandardCoil: The state of the left connection line is assigned to the corresponding BOOL
  variable (indicated by "***").
- NegatedCoil: The negated value of the left connection line state is assigned to the
  corresponding BOOL variable (indicated by "***"). If the left connection line state is
  Off, the corresponding variable is set to On, and if the left connection line state is
  On, the corresponding variable is set to Off.
- SetCoil: When the state of the left connection line becomes On, the corresponding BOOL
  variable (indicated by "***") is set to On and remains On until turned Off by the Reset
  coil.
- ResetCoil: When the state of the left connection line becomes On, the corresponding BOOL
  variable (indicated by "***") is set to Off and remains Off until turned On by the Set
  coil.
- RisingEdgeCoil: If the state of the left connection line changes from Off in the previous
  scan to On in the current scan, the value of the corresponding BOOL variable (indicated
  by "***") becomes On only during the current scan.
- FallingEdgeCoil: If the state of the left connection line changes from On in the previous
  scan to Off in the current scan, the value of the corresponding BOOL variable (indicated
  by "***") becomes On only during the current scan.

[Others]
- Inverter: The state of the left connection line is inverted and passed to the right
  connection line.
- FunctionBlock: Represents a function block.
- Variable: Represents the variable corresponding to the function.
- RisingEdge: Before detecting a positive transition, if the result of the previous
  operations changes from Off in the previous scan to On in the current scan, and the state
  of the left connection line is On, the state of the right connection line becomes On only
  during the current scan.
- FallingEdge: Before detecting a negative transition, if the result of the previous
  operations changes from On in the previous scan to Off in the current scan, and the state
  of the left connection line is On, the state of the right connection line becomes On only
  during the current scan.

```

Figure 9: System prompt for JSON.

G.3 System prompt for Code

```

You are a programming assistant specializing in generating Python code. Your task is to write
Python code that translates functional descriptions into equivalent PLC ladder logic and
represents the ladder logic as graphs using the NetworkX library. The natural language
instructions will describe the desired functionality. Your job is to:
1. Interpret the described functionality.
2. Translate it into equivalent ladder logic components (e.g., rungs, contacts, coils).
3. Implement this logic in Python code using NetworkX, representing the ladder logic as
   directed graphs.

### Requirements for Ladder Logic Representation:
- Nodes: Represent ladder logic elements such as inputs, outputs, and logic functions.
- Edges: Represent connections between these elements, indicating logical flow or sequence.

### ElementType of Nodes
Nodes perform differently based on their ElementType. The behavior for each ElementType is as
follows:
[Contact]
- NormallyOpen: When the state of the BOOL variable (indicated by "***") is On, the state of
  the left connection line is copied to the right connection line. Otherwise, the state of
  the right connection line is Off.
- NormallyClosed: When the state of the BOOL variable (indicated by "***") is Off, the state
  of the left connection line is copied to the right connection line. Otherwise, the state
  of the right connection line is Off.
- RisingEdgeContact: If the value of the BOOL variable (indicated by "***") changes from Off
  in the previous scan to On in the current scan, and the state of the left connection line
  is On, the state of the right connection line becomes On during the current scan.

```

- FallingEdgeContact: If the value of the BOOL variable (indicated by "***") changes from On in the previous scan to Off in the current scan, and the state of the left connection line is On, the state of the right connection line becomes On during the current scan.
- RisingEdgeNotContact: If the value of the BOOL variable (indicated by "***") changes from Off in the previous scan to On in the current scan, and the state of the left connection line is On, the state of the right connection line becomes Off during the current scan.
- FallingEdgeNotContact: If the value of the BOOL variable (indicated by "***") changes from On in the previous scan to Off in the current scan, and the state of the left connection line is On, the state of the right connection line becomes Off during the current scan.

[Coil]

- StandardCoil: The state of the left connection line is assigned to the corresponding BOOL variable (indicated by "***").
- NegatedCoil: The negated value of the left connection line state is assigned to the corresponding BOOL variable (indicated by "***"). If the left connection line state is Off, the corresponding variable is set to On, and if the left connection line state is On, the corresponding variable is set to Off.
- SetCoil: When the state of the left connection line becomes On, the corresponding BOOL variable (indicated by "***") is set to On and remains On until turned Off by the Reset coil.
- ResetCoil: When the state of the left connection line becomes On, the corresponding BOOL variable (indicated by "***") is set to Off and remains Off until turned On by the Set coil.
- RisingEdgeCoil: If the state of the left connection line changes from Off in the previous scan to On in the current scan, the value of the corresponding BOOL variable (indicated by "***") becomes On only during the current scan.
- FallingEdgeCoil: If the state of the left connection line changes from On in the previous scan to Off in the current scan, the value of the corresponding BOOL variable (indicated by "***") becomes On only during the current scan.

[Others]

- Inverter: The state of the Incoming edge is inverted and passed to the Outgoing edge.
- FunctionBlock: Represents a function block.
- Variable: Represents the variable corresponding to the function.

Guidelines

- Use the networkX library to define and manipulate the graph structure.
- Each rung in ladder logic must be represented as a separate directed graph.
- If the input describes multiple functionalities or rungs, your code should generate multiple graphs accordingly.

Figure 10: System prompt for Code.