GT-Fusion: Synergizing LLM Semantics, Topology, and Geometry for PPA Prediction and Optimization

Xinhua Lai[1,✉], He Liu[2], Miao Liu[1], Jungang Xu[1,*,✉], Xingquan Li[3,*,✉]
[1]University of Chinese Academy of Sciences, [2]Peking University, [3]Southeast University

***Abstract*—Existing AI-based PPA predictors often falter in reconciling the disparity between logical netlists and physical layouts, lacking a holistic view to balance conflicting design constraints. We propose GT-Fusion, a tri-modal framework establishing a Unified Multi-Objective Optimization Paradigm to bridge this gap. The framework synergizes an LLM-driven semantic encoder (Qwen3), a GNN-based topological extractor, and a UNet-based spatial analyzer via a Cross-Modal Injection Mechanism. Extensive experiments on Skywater 130nm benchmarks demonstrate that our tri-modal approach significantly outperforms both single-modal (Pure UNet) and bi-modal (UNet+GNN) baselines. Specifically, GT-Fusion reduces visual congestion prediction error (MAE) by 28.3% and path delay RMSE by 37.5%, proving the necessity of fusing semantic, topological, and geometric views. Furthermore, deployed within the iEDA open-source infrastructure, our model successfully guides placement optimization, delivering significant reductions in timing violations.**

***Index Terms*—VLSI, Physical Design, PPA Prediction, Large Language Models, Tri-modal Fusion, Closed-Loop Optimization**

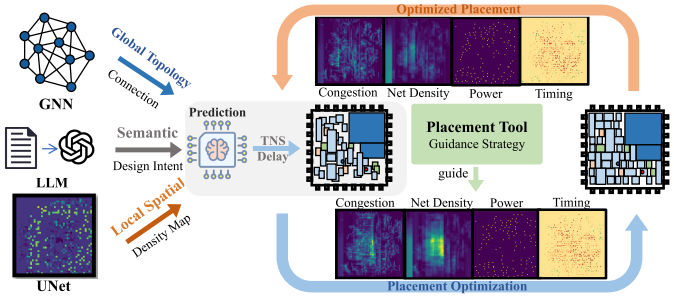


Fig. 1. **Motivation of GT-Fusion.** Current predictors rely on single modalities, missing the holistic view. Our framework fuses (1) **Global Topology** from GNNs, (2) **Semantic Design Intent** from LLMs, and (3) **Local Spatial Features** from UNets. This tri-modal synergy drives a **Prediction** module that feeds TNS/Delay and congestion maps into the **Placement Tool**, establishing a closed-loop **Placement Optimization** cycle.

I. Introduction

As Very Large Scale Integration (VLSI) technology advances into the nanometer era, the complexity of physical design continues to escalate. A critical challenge in modern Electronic Design Automation (EDA) flows is the widening correlation gap between the placement and routing stages. Power, Performance, and Area (PPA) metrics estimated during placement often deviate significantly from post-routing reality due to unmodeled parasitic effects and congestion-induced optimizations [1]. Consequently, accurate *Pre-routing PPA Prediction* has become a pivotal technology to enable "shift-left" design closure and reduce costly engineering change order (ECO) iterations [2].

Recent years have witnessed a paradigm shift from analytical models to data-driven machine learning solutions for PPA prediction. Early approaches utilized tree-based algorithms, such as Random Forest and XGBoost, to predict delay and power based on tabular features [3], [4]. While efficient, these methods struggle to capture the complex topological dependencies inherent in circuit netlists. To address this, Graph Neural Networks (GNNs) have emerged as a dominant methodology. For instance, He *et al.* combined GNNs with Transformers to capture both local neighbor effects and global path dependencies [5], while DeepSeq optimized GNN propagation specifically for sequential logic [6].

* Corresponding authors.

However, PPA metrics depend heavily on spatial characteristics like congestion, not just topology. While recent hybrid frameworks [7], [8] integrate Convolutional Neural Networks (CNNs) or Transformers to capture spatial features, they still treat designs purely as geometric graphs, often overlooking high-level **Design Intent**. Simultaneously, Large Language Models (LLMs) in EDA [9], [10] are predominantly restricted to generative tasks. They are rarely exploited as high-precision **feature encoders** capable of injecting semantic understanding into the numerical regression loop.

Consequently, methodologies remain fragmented across "modal islands": GNNs overlook layout; CV methods weaken logical representation; and LLMs are isolated from the optimization flow. Existing tools thus lack a unified paradigm to reconcile the trade-offs between semantic design intent (LLM), global topology (GNN), and local spatial characteristics (UNet).

To bridge this gap, we propose **GT-Fusion**, a novel tri-modal framework that synergizes LLM, GNN, and UNet to achieve closed-loop PPA optimization. As illustrated in Fig. 1, our design philosophy mimics a human engineer's cognitive process: comprehending functional intent via databooks (LLM), analyzing topological connectivity via schematics (GNN), and inspecting physical congestion via layout heatmaps (UNet). By aligning these heterogeneous views, our framework not only predicts potential violations but actively guides the placer to resolve them. The main contributions of this paper are summarized as follows:

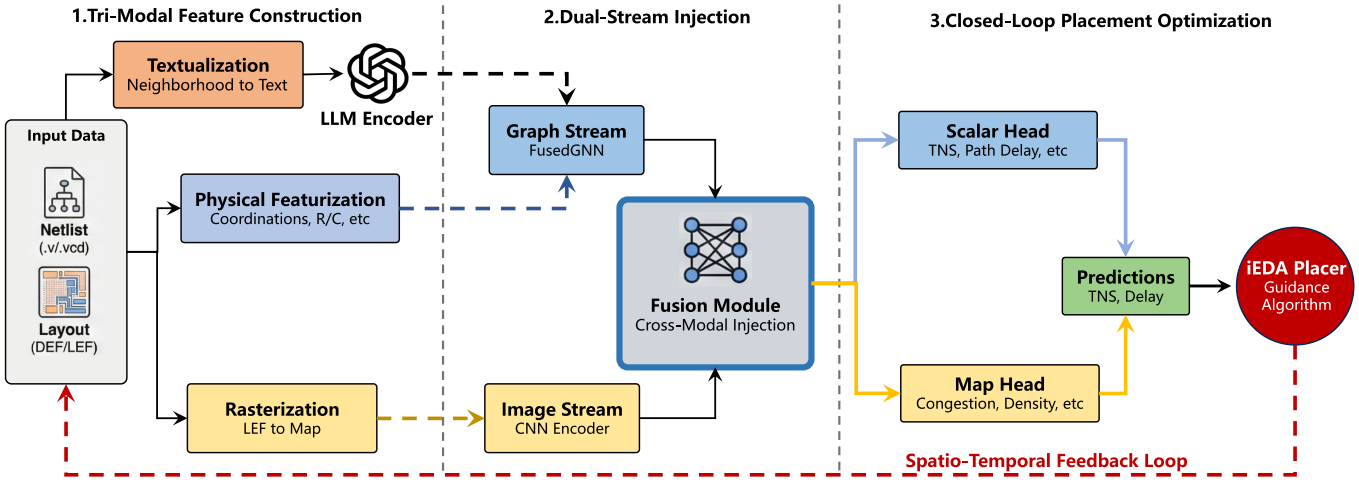


Fig. 2. **Overall Workflow of GT-Fusion.** The pipeline proceeds in three synergistic stages: (1) **Tri-Modal Feature Construction** (extracting semantic, topological, and spatial features); (2) **Dual-Stream Injection** (fusing embeddings via Cross-Modal Injection); and (3) **Closed-Loop Optimization** (guiding the iEDA placer with predicted PPA metrics).

- **Unified Multi-Objective Optimization Paradigm:** We propose a holistic framework that treats PPA closure as a global multi-objective problem. Unlike isolated predictors, our model acts as an adaptive guide, navigating the complex trade-off space (e.g., routability vs. performance) to dynamically prioritize critical objectives based on design characteristics.
- **Tri-Modal Architecture with Semantic Injection:** To power this framework, we develop GT-Fusion, synergizing LLM, GNN, and UNet. We innovatively repurpose Qwen3 as a high-precision *feature encoder* for design intent and introduce a *Cross-Modal Injection Mechanism* to align topological embeddings with spatial features in a unified latent space.
- **Closed-Loop Deployment in Open-Source EDA:** We bridge the gap to industrial application by integrating our model into the **iEDA** infrastructure. This establishes a practical closed-loop workflow where AI-generated guidance actively steers the placer, significantly outperforming single-modal baselines in reducing congestion and timing violations.

II. Problem Formulation

We formulate the pre-routing PPA prediction as a tri-modal regression task. Given a dataset $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}$, our goal is to learn a mapping function $\mathcal{F}_\Theta : \mathcal{X} \to \mathcal{Y}$ that projects circuit topology and semantics onto physical layout metrics.

Multi-modal Representation. The input $\mathcal{X}$ is represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node $v_i \in \mathcal{V}$ is characterized by its spatial coordinate $\mathbf{p}_i \in \mathbb{R}^2$ and a hybrid feature vector $\mathbf{h}_i$:

$$\mathbf{h}_i = \text{Concat}\left(\mathbf{f}_{phys}^{(i)}, \Phi_{LLM}(\text{text}_i)\right) \tag{1}$$

where $\mathbf{f}_{phys}^{(i)}$ denotes intrinsic physical attributes (e.g., cell size) and $\Phi_{LLM}(\cdot)$ denotes semantic embeddings extracted by an LLM to capture functional logic.

Learning Objective. The framework performs multi-task prediction to generate the target $\mathcal{Y}$, consisting of: (1) *Pixel-level Maps* $\hat{\mathbf{Y}}_{map} \in \mathbb{R}^{H\times W\times K}$ (e.g., congestion hotspots); and (2) *Design-level Scalars* $\hat{\mathbf{y}}_{scalar} \in \mathbb{R}^M$ (e.g., path delay, power). The model parameters Θ are optimized by minimizing the joint regression loss between the predictions and the Ground-Truth labels across the dataset.

III. Methodology

To address the challenges formulated in Section II, we propose GT-Fusion to bridge design intent and physical constraints. As illustrated in Fig. 2, the framework integrates three synergistic stages: (1) **Tri-Modal Feature Construction** for heterogeneous extraction; (2) **Dual-Stream Injection** for cross-modal fusion; and (3) **Closed-Loop Optimization** for predictive PPA guidance.

A. Hybrid Semantic-Physical Representation

To bridge the gap between implicit design intent and explicit physical constraints, we construct a hybrid representation scheme that integrates text, graph, and image modalities.

1) LLM-Driven Semantic Embedding: Traditional numerical features lack high-level functional context. We leverage **Qwen3** to extract semantic design intent. For each node v, we generate a textual description T_v of its local neighborhood (e.g., *"connected to u_adder..."*). This text is encoded into a high-dimensional vector h_{sem} via the LLM, capturing the logical functionality that pure statistics miss.

2) Heterogeneous Feature Encoders: The module processes two parallel streams to capture complementary physical views:

- **Graph Stream (Physical-Semantic Fusion):** We explicitly concatenate the LLM-derived semantic vector h_{sem} with physical node attributes h_{phy} (e.g., coordinates, fan-out, slew). The combined feature vector $h_{node} = [h_{sem}||h_{phy}]$ is then processed by a **FusedGNN** to capture global topology.

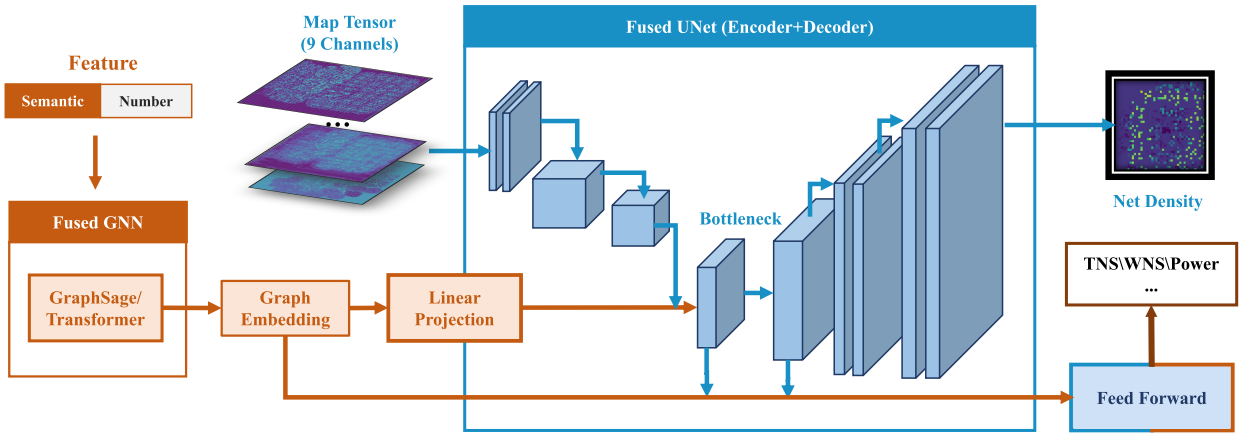


Fig. 3. **The Architecture of the GT-Fusion Model.** Key innovation lies in the *Cross-Modal Injection Module*, where the global graph embedding (H_{graph}) is projected and broadcasted to fuse with the visual bottleneck (H_{image}), aligning logical semantics with spatial reconstruction.

- **Image Stream (Spatial View):** Concurrently, the placement layout is rasterized into multi-channel feature maps (e.g., Cell Density, Pin Density). A CNN-based encoder processes these maps to extract local spatial patterns and congestion risks.

B. The GT-Fusion Architecture

As shown in Fig. 3, GT-Fusion integrates the topological semantics of the circuit netlist into the geometric representation of the physical layout. The framework consists of a Graph Encoder (GNN), a Layout Encoder (UNet Encoder), and a novel Cross-Modal Injection Mechanism that aligns these two heterogeneous modalities in the latent space.

1) Latent Space Alignment & Injection: A critical challenge lies in bridging the dimensional gap between the graph embedding $v_{graph} \in \mathbb{R}^{D_g}$ and the layout feature map $F_{img} \in \mathbb{R}^{C\times H\times W}$. To achieve this efficiently without introducing excessive parameters, we propose a lightweight **Additive Semantic Injection** strategy.

First, the graph embedding is projected via a linear transformation $\mathcal{W}_p$ to align with the visual latent space:

$$v'_{graph} = \sigma(\mathcal{W}_p \cdot v_{graph} + b_p) \tag{2}$$

where $v'_{graph} \in \mathbb{R}^C$ matches the channel depth. Subsequently, this global context vector is spatially broadcasted and injected into the visual stream via element-wise addition:

$$F_{fused} = F_{img} \oplus \text{Tile}(v'_{graph}) \tag{3}$$

This mechanism functions as a **global semantic bias**. By shifting the feature manifold of the layout representation, it effectively trains the decoder on the circuit topology, guiding the reconstruction of congestion hotspots while preserving the gradient flow stability inherent to residual connections.

2) Decoupled Multi-Task Decoders: To achieve holistic PPA prediction, the architecture bifurcates into two task-specific heads, decoupling spatial reconstruction from scalar regression: **Pixel Reconstruction Head.** This branch mirrors the expansion path of a standard UNet. It utilizes the spatially fused features through a series of up-sampling and convolution blocks to progressively recover fine-grained details. The output is a high-resolution density map $\hat{Y}_{map} \in \mathbb{R}^{H_{out}\times W_{out}}$ (e.g., Congestion).

Scalar Regression Head. Parallel to the pixel decoder, this head extracts chip-level metrics. Unlike the spatial path, we employ an **Explicit Concatenation** strategy to maximize information retention. We extract the global visual embedding $v_{img} = \text{GAP}(F_{enc})$ from the bottleneck and concatenate it directly with the original graph embedding v_{graph}:

$$\hat{Y}_{scalar} = \text{MLP}\left(\text{Concat}[v_{img}, v_{graph}]\right) \tag{4}$$

This design allows the regressor to leverage the distinct statistical distributions of both modalities for accurate metric prediction (e.g., TNS, Power).

C. Physics-Aware Learning Strategy

1) Global Statistics Normalization: To address the extreme dynamic ranges of EDA data (e.g., capacitance 10^{-15} vs. coordinates 10^3), we implement a strict **Global Statistics Normalization** strategy derived exclusively from the training split $\mathcal{D}_{train}$. **Graph Standardization.** For netlist features, we apply distinct standard scalers (z-score) for node, edge, and graph-level attributes using global mean μ and std σ to ensure balanced gradient contributions. **Layout Normalization.** Unlike Instance Normalization which destroys relative density information, we adopt **Channel-wise Global Normalization**. Statistics are computed across the entire training dataset to preserve critical relative magnitudes between different chip designs. During inference, an **Inverse Normalization** module restores predictions to physical units (e.g., ns, mW) for EDA tool compatibility.

2) Joint Optimization: The model is trained end-to-end using a hybrid objective that balances spatial reconstruction and global regression. We employ a weighted sum of Mean Squared Error (MSE) losses:

$$\mathcal{L}_{total} = \lambda_{map} \underbrace{\|\mathbf{Y}_{map} - \hat{\mathbf{Y}}_{map}\|_2^2}_{\text{Spatial Reconstruction}} + \lambda_{scalar} \underbrace{\|\mathbf{y}_{scalar} - \hat{\mathbf{y}}_{scalar}\|_2^2}_{\text{Global Regression}} \tag{5}$$

TABLE I

QUANTITATIVE PERFORMANCE COMPARISON ON THE UNSEEN TEST SET. THE PROPOSED FULL MODEL DEMONSTRATES SUPERIOR FIDELITY IN VISUAL CONGESTION PREDICTION AND CRITICAL TIMING METRICS (CLOCK FREQUENCY AND PATH DELAYS).

Metric Type	Baseline (Pure UNet)				Ablation (UNet + GNN)				Ours (Full Model)			
	RMSE↓	MAE↓	PSNR↑	IoU↑	RMSE↓	MAE↓	PSNR↑	IoU↑	RMSE↓	MAE↓	PSNR↑	IoU↑
Visual Prediction Task												
Congestion Map	0.3190	0.1273	18.09	0.3244	0.2906	0.0982	19.21	0.3664	**0.2617**	**0.0913**	**21.30**	**0.3808**
Scalar Prediction Tasks												
Clock Freq (MHz)	421.66	254.81	-		392.82	235.45	-		**378.08**	**209.06**	-	
TNS (ns)	**56.15**	**46.21**	-		198.37	103.00	-		451.88	240.62	-	
Path Delays (ns)	1.1031	0.6888	-		0.7939	0.6114	-		**0.6895**	**0.5578**	-	
Leakage Pwr (mW)	0.7640	0.5980	-		**0.5199**	**0.4864**	-		0.8776	0.6806	-	
Internal Pwr (mW)	0.7520	0.5936	-		**0.5470**	**0.4854**	-		1.1447	0.8271	-	
Switch Pwr (mW)	0.7632	**0.6024**	-		**0.7629**	0.6172	-		1.2018	0.8207	-	

where $\lambda_{map} = 1.0$ and $\lambda_{scalar} = 0.25$. This weighting scheme prioritizes the structural fidelity of the congestion map while using scalar regression as an auxiliary regularization task. We optimize using Adam with **Gradient Accumulation** to simulate larger batch sizes, stabilizing convergence under high-resolution memory constraints.

D. AI-Guided Closed-Loop Optimization

To bridge the gap between early-stage placement and sign-off timing, we propose a holistic closed-loop strategy that injects predicted Total Negative Slack (TNS) and Worst Negative Slack (WNS) directly into the placer's gradient descent loop. We introduce a **Hybrid Weighting Scheme** where the net weight update is modulated by an scaling factor Γ_n:

$$w_n^{(k+1)} = w_n^{(k)} + \Delta w_n^{base} \cdot \Gamma_n(\mathcal{G}_{pred}) \tag{6}$$

Specifically, Γ_n balances global convergence with local criticality via a two-stage mechanism:

1. Global Aggressiveness (α_{TNS}): We use logarithmic scaling of the predicted TNS to set the baseline optimization intensity, ensuring numerical stability while reflecting design closure difficulty:

$$\alpha_{TNS} = 1 + \log_{10}(1 + |TNS_{pred}|) \tag{7}$$

2. Targeted Criticality (WNS Factor): Since exact Slack depends on clock constraints unavailable during global placement, we use the **Predicted Longest Path Delay** (D_{pred}) as a proxy for WNS severity. To avoid wirelength bloat on non-critical paths, we apply a targeted penalty derived from D_{pred} only to nets with high physical centrality ($\mu_n > 0.5$):

$$\Gamma_n = \begin{cases} \alpha_{\text{TNS}} \left[1 + (\mu_n - 0.5)(1 + 0.5 D_{\text{pred}})\right], & \text{if } \mu_n > 0.5 \\ \alpha_{\text{TNS}}, & \text{otherwise} \end{cases} \tag{8}$$

By using the positive delay value D_{pred}, the penalty scales proportionally with the path length, aggressively pulling cells on long paths closer while maintaining a conservative strategy for the rest of the design.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

- **Datasets and Baselines:** Based on the AiEDA and iEDA platform [11], [12], we evaluate three models: (1) a baseline **Pure UNet**, (2) an ablation **GNN+UNet**, and (3) our **Tri-modal Framework**. The model is further integrated into the *iPL* tool (iEDA) to validate closed-loop layout optimization.
- **Evaluation Metrics:** We assess pixel-level fidelity via **RMSE, MAE**, and **Peak Signal-to-Noise Ratio (PSNR)**. For hotspot localization, we calculate **Intersection over Union (IoU)** on critical regions defined by the **top 3%** intensity threshold.
- **Implementation:** Experiments are conducted on a server with 96 Intel Xeon 8168 CPUs, 1.5TB RAM, and 8 NVIDIA Tesla V100 GPUs. The framework is implemented using PyTorch 2.4.1 and CUDA 12.2.

B. Metrics Prediction

Table I presents the quantitative comparison of our proposed Full Model against the Baseline (Pure UNet) and the Ablation model (UNet + GNN) on the unseen test set. The results demonstrate that integrating semantic knowledge from LLMs with structural graph features yields significant performance gains in the most critical placement tasks.

Dominance in Visual Congestion Prediction. In the congestion map prediction task, the Full Model achieves a "grand slam" by outperforming all baselines across every evaluation metric. Most notably, it attains the highest **IoU of 0.3808** (a **17.4%** improvement over the Baseline) and the highest **PSNR of 21.30 dB**. Unlike the Ablation model, which improves structural consistency but struggles with pixel-level precision, the Full Model also achieves the lowest **RMSE (0.2617)** and **MAE (0.0913)**. This indicates that the semantic embeddings enable the predictor to not only localize congestion hotspots accurately (High IoU) but also reconstruct the global congestion distribution with high fidelity (Low RMSE).

Superiority in Timing Estimation. For scalar prediction tasks, the Full Model exhibits exceptional accuracy in timing-related metrics, which are heavily dependent on logic functionality. It reduces the MAE of *Clock Frequency* to **209.06**

TABLE II

QUANTITATIVE EVALUATION OF CLOSED-LOOP OPTIMIZATION. THE RESULTS DEMONSTRATE THE MODEL'S ADAPTABILITY: IT PRIORITIZES ROUTABILITY (REDUCING *Max Overflow* BY 16.1%) IN THE CONGESTION-HEAVY `salsa20`, WHILE OPTIMIZING TIMING (IMPROVING *WNS* BY 3.3%) IN THE TIMING-CRITICAL `s9234`.

Metric	**s713** (Small)			**s9234** (Timing-Critical)			**salsa20** (Congestion-Heavy)		
	Base	**AI**	**Imp (%)**	**Base**	**AI**	**Imp (%)**	**Base**	**AI**	**Imp (%)**
Global Routing & Resources (Lower is Better)									
Total Overflow	159	**157**	**-1.26%** ↓	1659	1875	+13.02% ↑	19362	**17647**	**-8.86%** ↓
Max Overflow	10	10	+0.00% →	24	28	+16.67% ↑	31	**26**	**-16.13%** ↓
Wirelength (10^6 DBU)	2.898	**2.890**	**-0.28%** ↓	18.561	**18.208**	**-1.90%** ↓	150.238	**149.744**	**-0.33%** ↓
Timing, Power & Performance (Higher/Closer to 0 is Better)									
WNS (ns)	-0.1389	**-0.1253**	**+9.76%** ↑	-6.0536	**-5.8518**	**+3.33%** ↑	0.0000	0.0000	+0.00% ↑
TNS (ns)	-0.0791	**-0.0791**	**+0.04%** ↑	-0.2760	**-0.2693**	**+2.40%** ↑	**1.8774**	1.8729	-0.24% ↓
Freq (MHz)	633.27	**633.28**	**+0.00%** ↑	563.08	**565.19**	**+0.37%** ↑	**1606.1**	1594.7	-0.71% ↓
Dynamic Pwr (e-4)	**8.5015**	8.5212	+0.23% ↑	74.4368	**74.4124**	**-0.03%** ↓	504.36	**504.35**	-0.00% →
Map Distribution Metrics									
Net Density	2.7208	**2.5331**	**-6.90%** ↓	3.7775	**2.4228**	**-35.86%** ↓	1.9659	**1.7793**	**-9.49%** ↓
Cell Density	0.3386	**0.3248**	**-4.09%** ↓	0.2657	**0.2631**	**-0.98%** ↓	0.2343	**0.2293**	**-2.14%** ↓
Congestion (e-4)	5.2532	**5.1771**	**-1.45%** ↓	4.9583	**3.8731**	**-21.89%** ↓	2.1001	2.1274	+1.30% ↑
Power (e-4)	7.6451	**7.6433**	**-0.02%** ↓	67.3316	**67.3176**	**-0.02%** ↓	476.22	**475.25**	**-0.21%** ↓

MHz and the MAE of *Top-100 Path Delays* to **0.5578 ns**, representing a **19.0%** error reduction compared to the baseline. This confirms that the LLM-extracted features successfully capture the logic depth and timing criticality of the netlist, providing reliable guidance for timing-driven placement.

Performance Trade-offs. While the ablation model (UNet + GNN) shows marginally better performance in power estimation metrics (e.g., Leakage and Internal Power), the Full Model strategically prioritizes the accuracy of routability (congestion) and performance (timing). Given that congestion overflow and timing violations are the primary bottlenecks preventing design closure, the substantial gains in these areas outweigh the minor degradation in power prediction accuracy. The results validate our design philosophy: the Full Model is a specialized expert for solving the hardest constraints in physical design.

C. Closed-Loop Placement Optimization in iEDA

Table II presents a quantitative comparison between the baseline and our GT-Fusion framework across three representative benchmarks (`s713`, `s9234`, `salsa20`). The results demonstrate the model's adaptability in navigating the trade-off space between routability, timing, and power constraints.

Routability-Driven Optimization (Congestion-Heavy). On the congestion-dominated design `salsa20`, GT-Fusion explicitly prioritizes routability. It reduces *Total Overflow* by **8.86%** and *Max Overflow* by **16.13%**, effectively eliminating critical hotspots by optimizing resource distribution. This is further evidenced by a **9.49%** drop in *Net Density*. Although the global average congestion shows a marginal increase (+1.30%), this reflects a deliberate strategy of "logic spreading"—distributing cells from peak-density regions to sparse areas to resolve local violations without degrading timing closure.

Performance-Driven Optimization (Timing-Critical). For the timing-critical design `s9234`, the framework adaptively shifts focus to path optimization. It improves *WNS* by **3.33%** and *TNS* by **2.40%**, while reducing *Wirelength* by **1.90%**. Notably, *Net Density* drops by a substantial **35.86%**, indicating the creation of essential routing channels for critical nets. The increase in *Total Overflow* (+13.02%) represents a calculated trade-off: the placer implicitly relaxes constraints in non-critical regions (accepting minor detours) to maximize slack on critical paths.

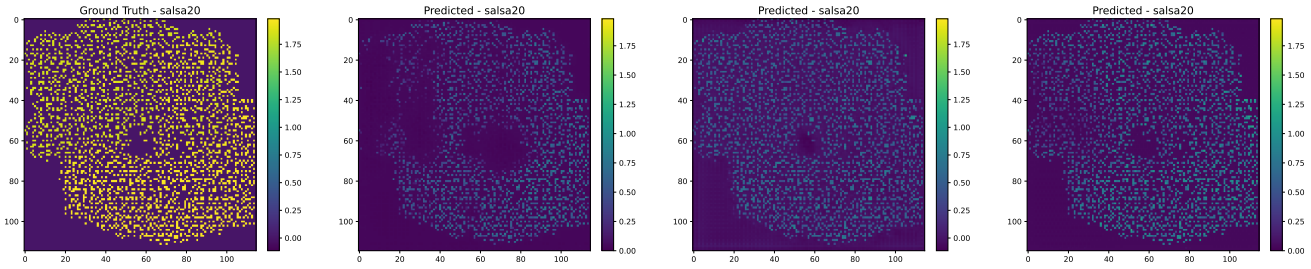
Balanced Optimization (General). On the smaller design `s713`, our approach achieves comprehensive gains, reducing *Total Overflow* by **1.26%** while simultaneously improving *WNS* by **9.76%**. The consistent reductions in *Power Map* sums across all designs further validate that our guidance effectively steers cells toward lower-density, power-efficient regions.

In summary, GT-Fusion acts as an adaptive optimization agent: aggressively mitigating hotspots in routability-limited scenarios while maximizing slack and wirelength efficiency in timing-limited designs.

D. Ablation Studies and Visual Analysis

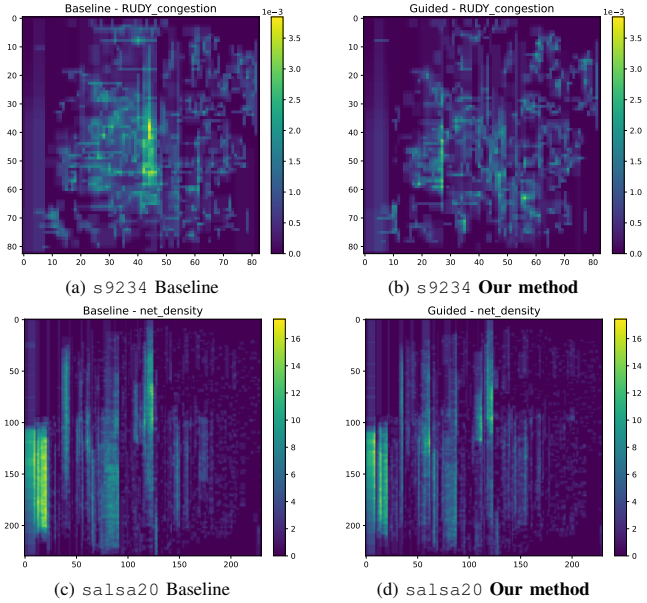
To provide a holistic validation of our framework, we conduct ablation studies from two complementary perspectives: (1) **Model Component Analysis**, visualizing how each module contributes to prediction fidelity; (2) **Guidance Effectiveness Analysis**, visualizing how the predicted maps translate into tangible improvements in the final physical layout.

1) Visualizing Model Component Contributions: Fig. 4 visually decomposes the contribution of GNN and LLM modules by comparing the net density map predictions against the Ground-Truth. As shown in Fig. 4(b), the **Pure UNet** captures only partial density trends; notably, it fails to predict the large high-density regions in the center and top-left, resulting in a loss of critical details. In contrast, **incorporating the GNN** (Fig. 4(c)) successfully recovers the complete density distribution trend, demonstrating the GNN's superior capability in capturing topological dependencies. Finally, the **Full Model** (Fig. 4(d)) generates the most structurally faithful prediction. By injecting LLM-based semantic embeddings into the graph encoder, it not only reconstructs the complete density distribution but also achieves the highest fidelity to the Ground-Truth.



(a) Ground-Truth (b) Baseline (Pure UNet) (c) Ablation (w/o LLM) (d) **Ours (Full Model)**

Fig. 4. **Visual Ablation of Prediction Fidelity.** Comparison on `salsa20`. While the Baseline (b) is blurry and the ablation model (c) is noisy, our Full Model (d) accurately reconstructs the high-density hotspots observed in the Ground-Truth (a).



(a) `s9234` Baseline (b) `s9234` **Our method**

(c) `salsa20` Baseline (d) `salsa20` **Our method**

Fig. 5. **Visual Impact on Physical Layout.** Top row: For the timing-critical `s9234`, the AI guide (b) promotes logic spreading to improve WNS. Bottom row: For the congestion-heavy `salsa20`, the AI guide (d) eliminates the central hotspots seen in the baseline (c) to resolve overflow.

2) Visualizing GT-Fusion Guidance on Physical Layout: Fig. 5 demonstrates the impact of GT-Fusion on placement quality by distinguishing between congestion mitigation and density optimization.

Case Study 1: Alleviating Congestion for Timing Closure (`s9234`). For this timing-critical design, the baseline congestion map (Fig. 5(a)) exposes severe hotspots that force signal detours. In contrast, our optimized layout (Fig. 5(b)) achieves a more uniform distribution. This **21.89%** reduction in peak congestion creates essential routing resources, allowing the router to fix critical paths with straighter connections. This improvement translates to a **3.33%** gain in WNS and a **74.4 mW** reduction in dynamic power.

Case Study 2: Homogenizing Density for Routability (`salsa20`). In the Net Density comparison (Fig. 5(c) vs. (d)), the baseline suffers from extreme local accumulation (bright yellow regions), correlating with routing failures in Table II. GT-Fusion successfully diffuses these clusters by leveraging global semantic-topological foresight. By reducing peak Net Density by **9.49%** (1.97 $\rightarrow$ 1.78), our method eliminates the root cause of shorts, ensuring a DRC-clean layout without sacrificing area.

V. Conclusion

In this paper, we presented **GT-Fusion**, a unified framework designed to break the "modal islands" in physical design. By synergizing LLM semantics, GNN topology, and UNet geometry, our approach bridges the gap between abstract design intent and physical constraints. Unlike isolated predictors, we repurpose the LLM as a feature encoder to power a cross-modal injection mechanism. Experimental results on Skywater 130nm benchmarks demonstrate the efficacy of this fusion, reducing congestion prediction error (MAE) by **28.3%** and path delay RMSE by **37.5%** compared to unimodal baselines. Furthermore, by establishing a practical **closed-loop optimization flow** within the **iEDA** infrastructure, this work marks a pivotal step toward autonomous, semantics-aware chip design.

References

[1] Z. He, Y. Pu, H. Wu *et al.*, "Large language models for eda: Future or mirage?" *ACM TODAES*, vol. 30, no. 6, pp. 1–53, 2025.
[2] D. Chen, V. Ganesh, W. Li *et al.*, "Report for nsf workshop on ai for electronic design automation," *arXiv:2601.14541*, 2026.
[3] Y. Attaoui, M. Chentouf, Z. E. A. A. Ismaili, and A. El Mourabit, "Enhancing cell delay accuracy in post-placed netlists using ensemble tree-based algorithms," *Integration*, vol. 97, p. 102193, 2024.
[4] S. Satapathy and D. S. Banerjee, "Dapp: Delay aware power prediction," in *Proc. ISQED*, 2025.
[5] G. He, W. Ding, Y. Ye *et al.*, "An optimization-aware pre-routing timing prediction framework based on heterogeneous graph learning," in *Proc. ASP-DAC*, 2024.
[6] S. Khan, Z. Shi, M. Li, and Q. Xu, "Deepseq: Deep sequential circuit learning," in *Proc. DATE*, 2024.
[7] H. Liu, Z. Zeng, S. Tao *et al.*, "Aitpo: Kan-unet heterogeneous network for timing prediction and optimization at global routing," *ACM TODAES*, vol. 31, no. 3, pp. 1–28, 2025.
[8] P. Cao, Y. Qin, G. He, W. Ding, X. Cheng, Z. Zhang, and Y. Ye, "An optimization-aware pre-routing timing prediction framework based on multi-modal learning," *IEEE TCAD*, vol. 44, no. 10, pp. 3896–3909, 2025.
[9] J. Pan, G. Zhou, C. Chang *et al.*, "A survey of research in large language models for electronic design automation," *ACM TODAES*, vol. 30, no. 3, pp. 1–21, 2025.
[10] W. Li, Y. Zou, C. Ellis *et al.*, "Bridges: Bridging graph modality and large language models within eda tasks," *arXiv:2504.05180*, 2025.
[11] Y. Qiu, Z. Huang, S. Tao, H. Zhang, W. Li, X. Lai, R. Wang, W. Wang, and X. Li, "Aieda: An open-source ai-aided design library for design-to-vector," *IEEE TCAD*, 2025.
[12] X. Li, Z. Huang, S. Tao *et al.*, "ieda: An open-source infrastructure of eda," in *Proc. ASP-DAC*, 2024, pp. 77–82.