Network and Distributed System Security (NDSS) Symposium 2026

ViGText: Deepfake Image Detection with Vision-Language Model Explanations and Graph Neural Networks

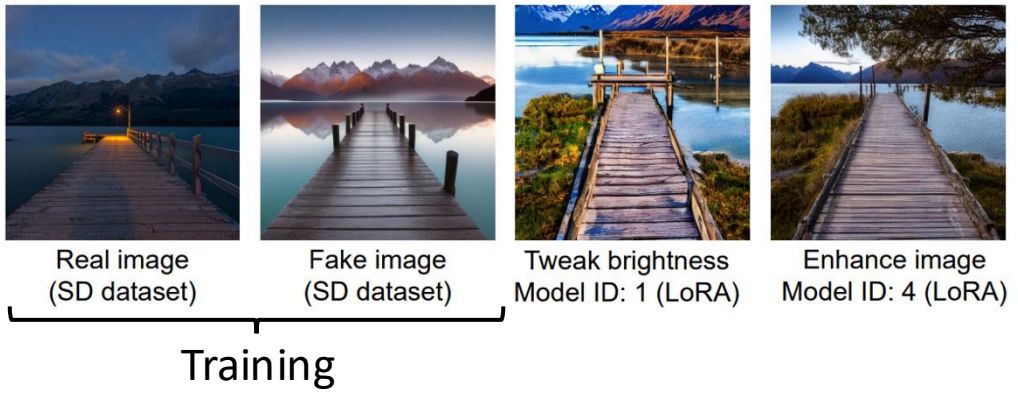
Ahmad ALBarqawi*, Mahmoud Nazzal§, Issa Khalil‖,
Abdallah Khreishah*, and NhatHai Phan*

*New Jersey Institute of Technology, Newark, NJ, USA
§Old Dominion University, Norfolk, VA, USA
‖Qatar Computing Research Institute (QCRI), HBKU, Doha, Qatar

Presented by:
Ahmad ALBarqawi
Contact: aka87@njit.edu
February 25, 2026 — San Diego, CA, USA

Background: Rethinking Deepfake Detection

- **AI-generated images are increasingly indistinguishable from real ones.**
- **The Core Challenge**
  - **Attackers are empowered [1]**
    - Fine-tuned variants (LoRA, FM) evade detectors
    - Subtle foundation model optimization attacks bypasses detection
  - **Naïve Defenders (DE-FAKE [2])**
    - Uses generic captions
    - Simple image–text embedding concatenation



(a) Source image (b) Manipulated image

Figures adapted from [1]

LoRA: Low Rank Adaptation
FM: Full Model, SD: Stable Diffusion

ViGText: Detailed Explanations

- VLMs → Human-like forensic explanations (e.g., lighting inconsistencies, texture artifacts)
- Patch-level descriptions aligned with specific regions
- Textual insights serve as contextual features

A Sample Textual Explanation

{B3,B4}: The window blinds have uneven spacing, and the light passing through does not align properly with the individual slats, which suggests an error in rendering light and shadows. {D1,D2}: The oven appears to have a distorted handle, and the reflection and shadow around it don't conform to the expected perspective and lighting. {D3}: The drawer underneath the stove has irregular handles that are asymmetrical, which is not typical for kitchen design and could be an oversight by the AI.

VLM: Vision Language Model

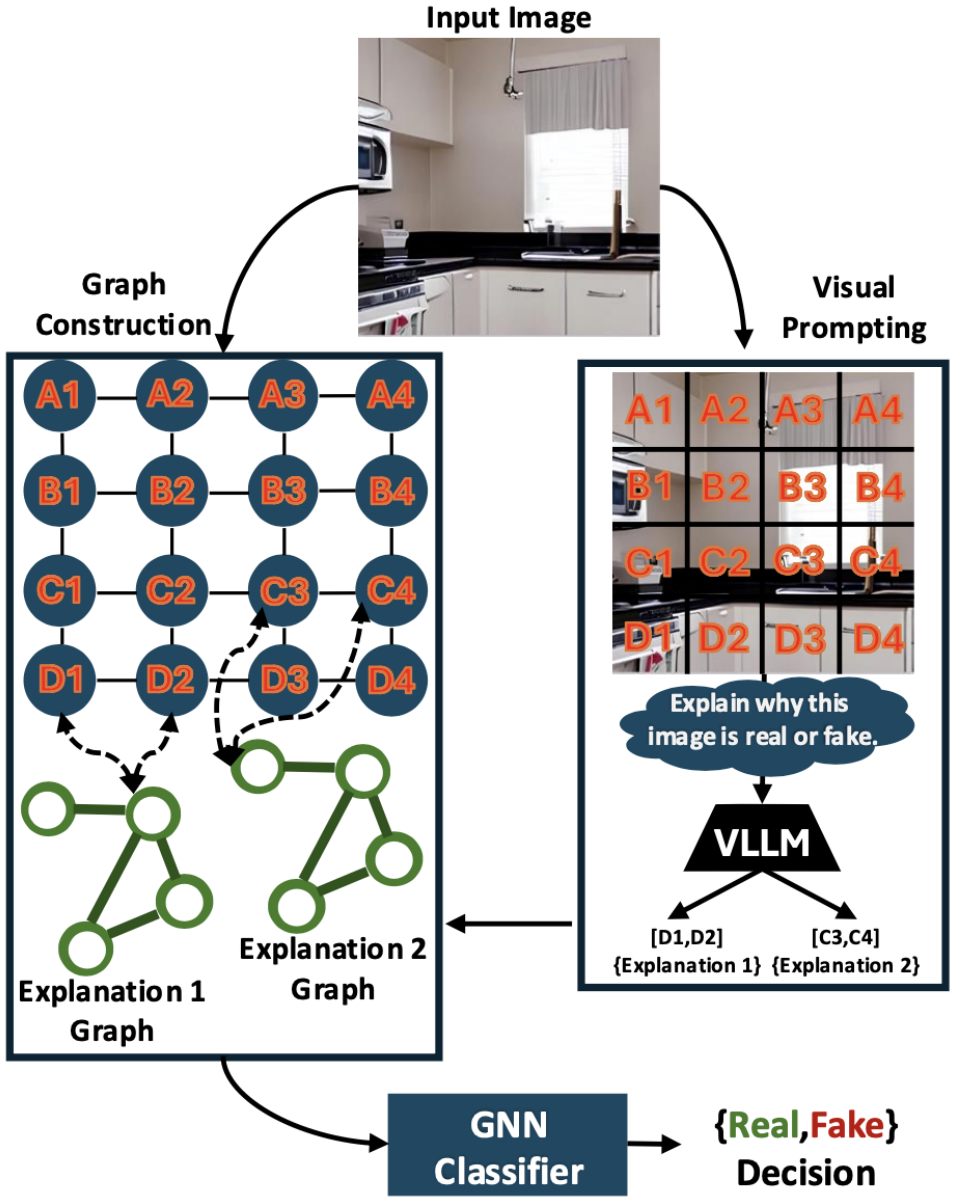
ViGText: Why Is Better Integration Needed?

- Naive text–image integration underutilizes rich explanations
- DE-FAKE [2]: marginal gains with VLM explanations
- Motivates context-aware multimodal integration

	Accuracy	Recall	Precision	F1
DE-FAKE w/Explanations	90.00	91.20	89.00	90.10
ViGText	**99.25**	**99.80**	**98.52**	**99.26**

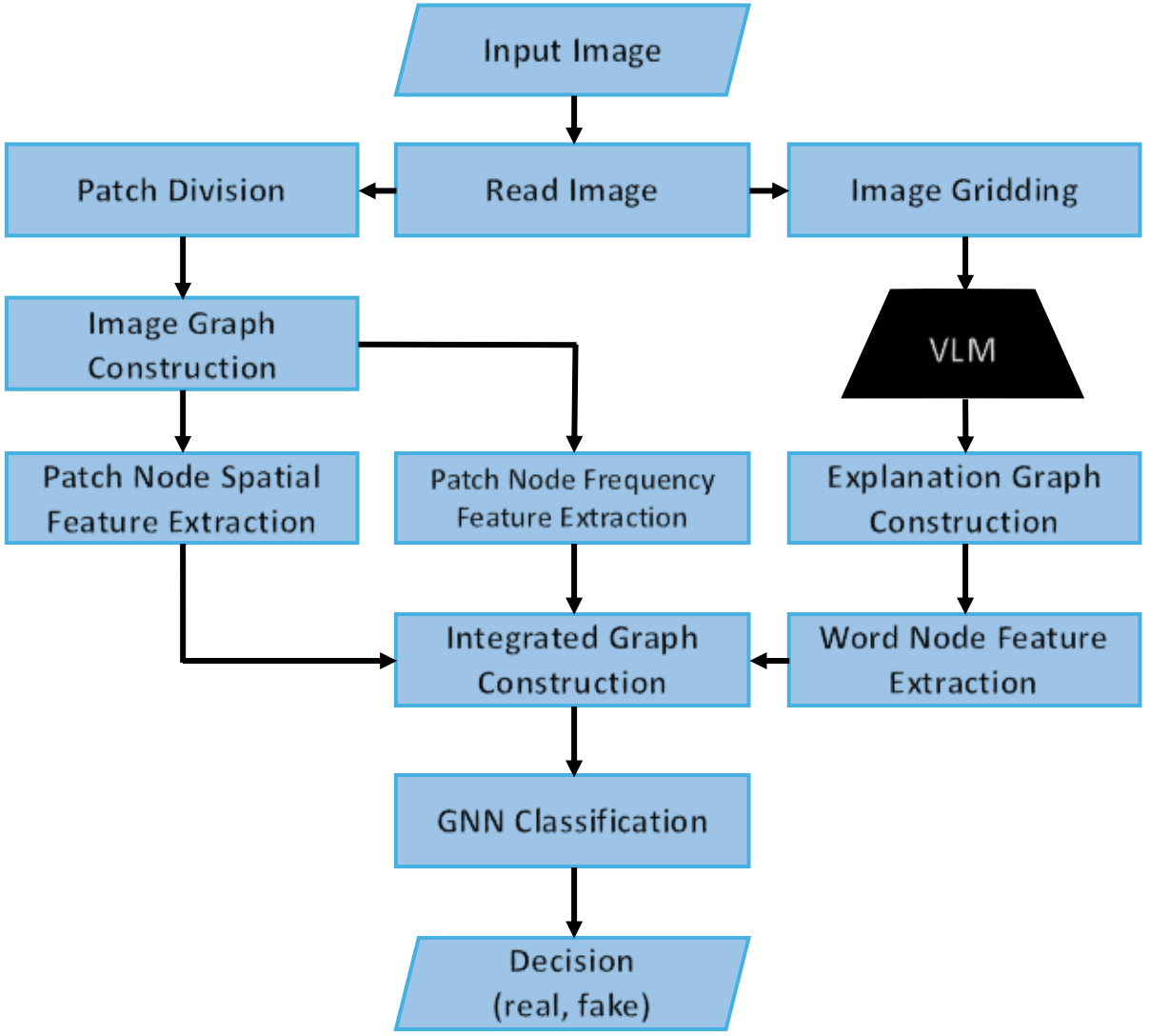
ViGText: Patch-Level Reasoning via Vision-Language Graphs

- VLMs → generate patch-level explanations
- Region–text integration enables context-aware reasoning
- image–text graphs → relational inference with Graph Neural Network (GNN)



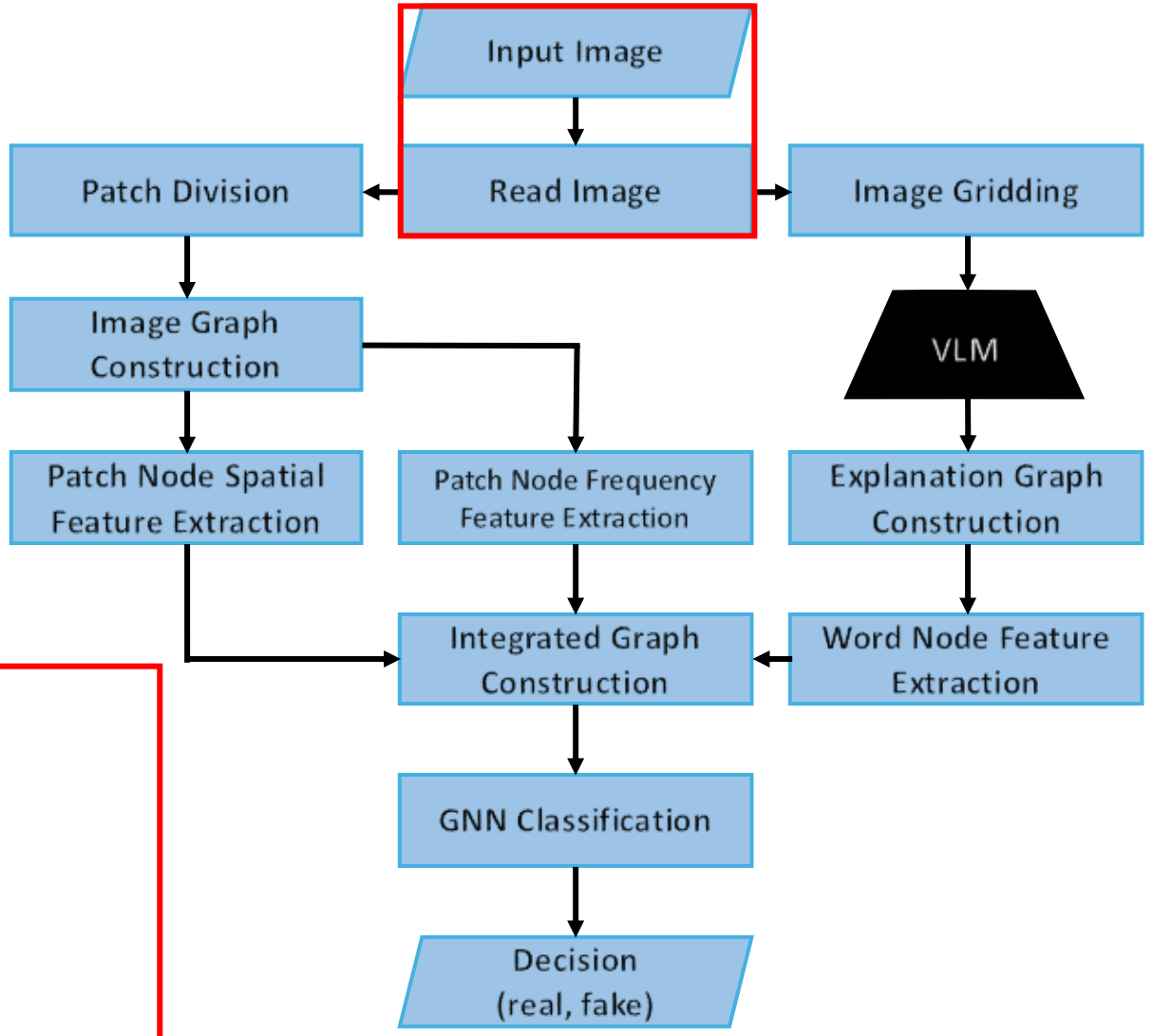
ViGText: The Pipeline

- Integrated graphs → image graph + explanation graph
- Patch-wise multimodal feature fusion (spatial, frequency, text)
- GNN → deepfake detection decision



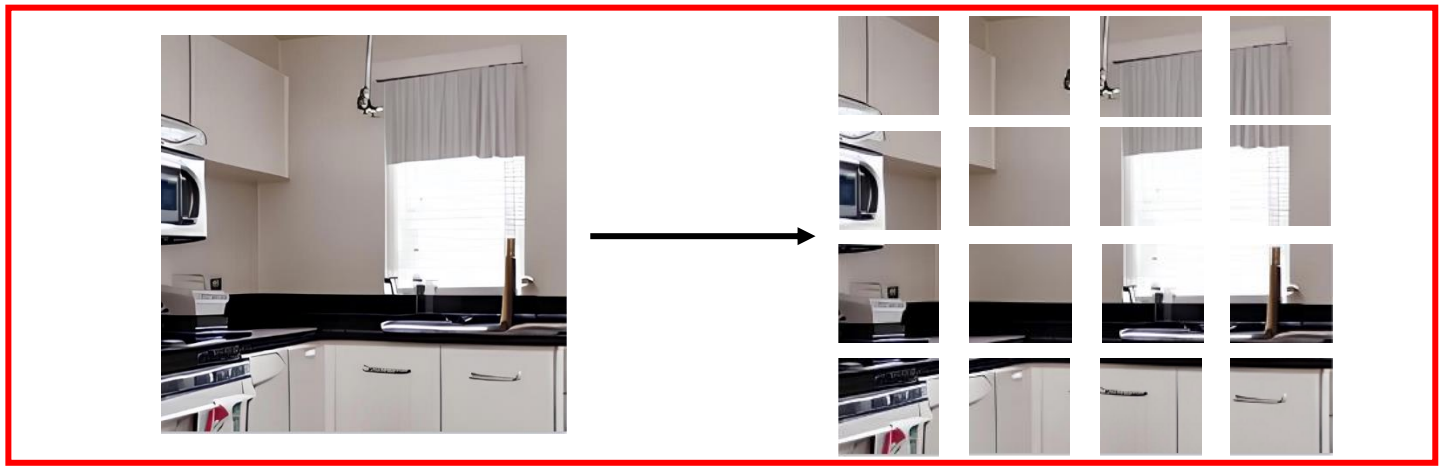
ViGText: The Pipeline

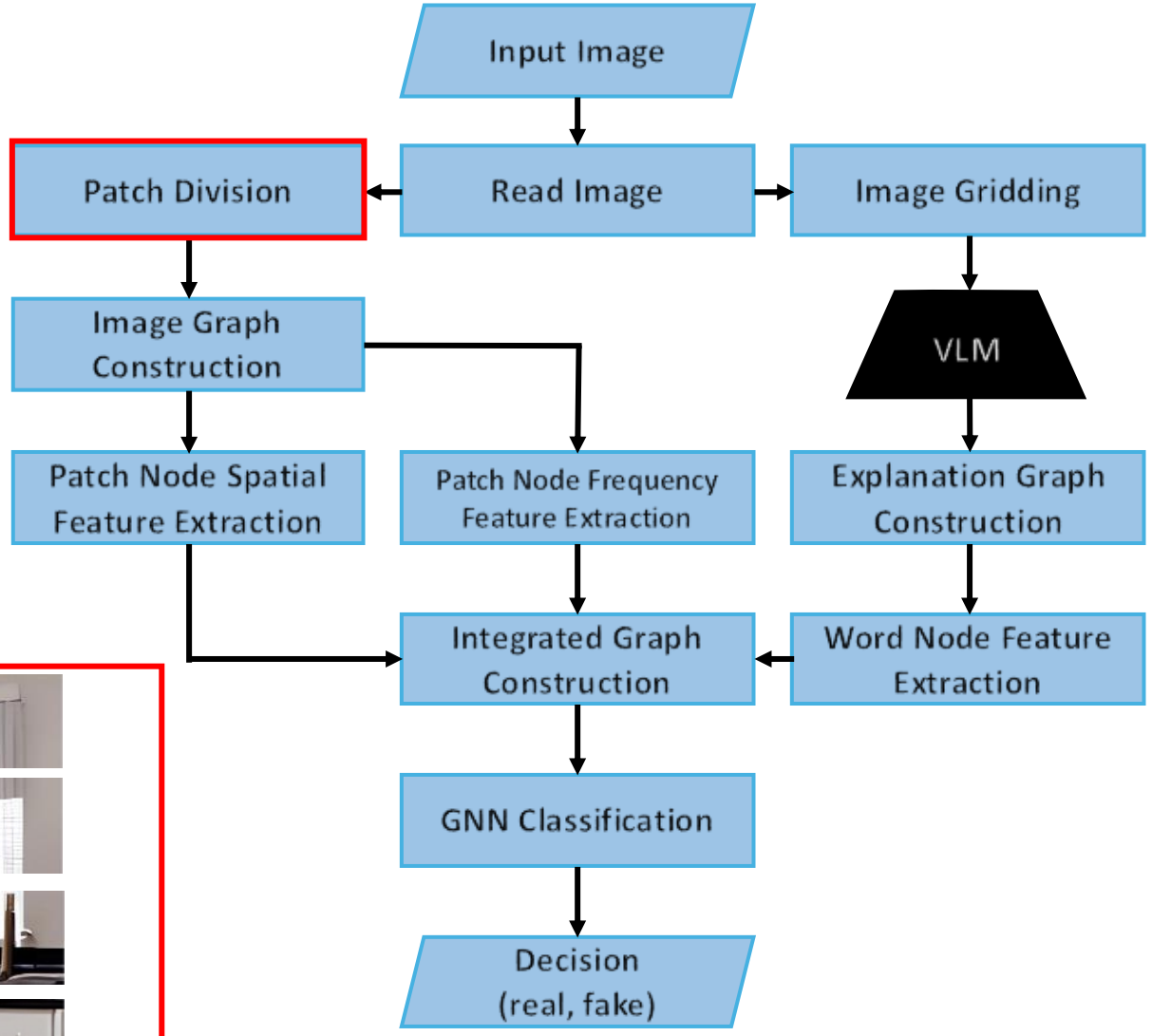
- The ViGText pipeline starts with any image:
  - Generated/Real
  - Any Resolution
  - Any Aspect Ratio



ViGText: The Pipeline

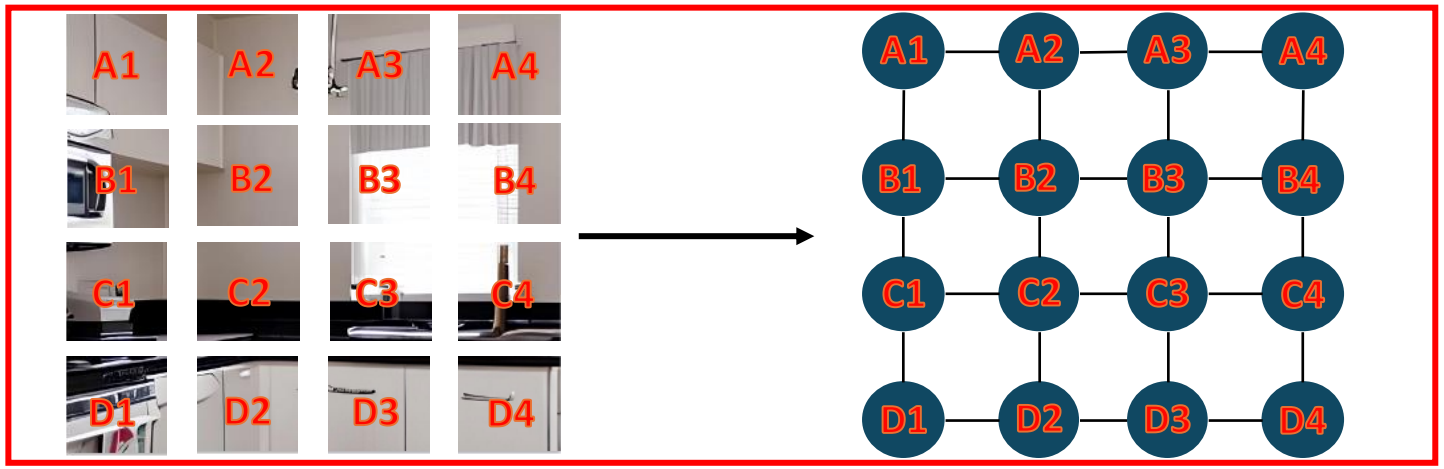
- Image divided into 16 uniform patches
- Prepares patch-level features for graph construction

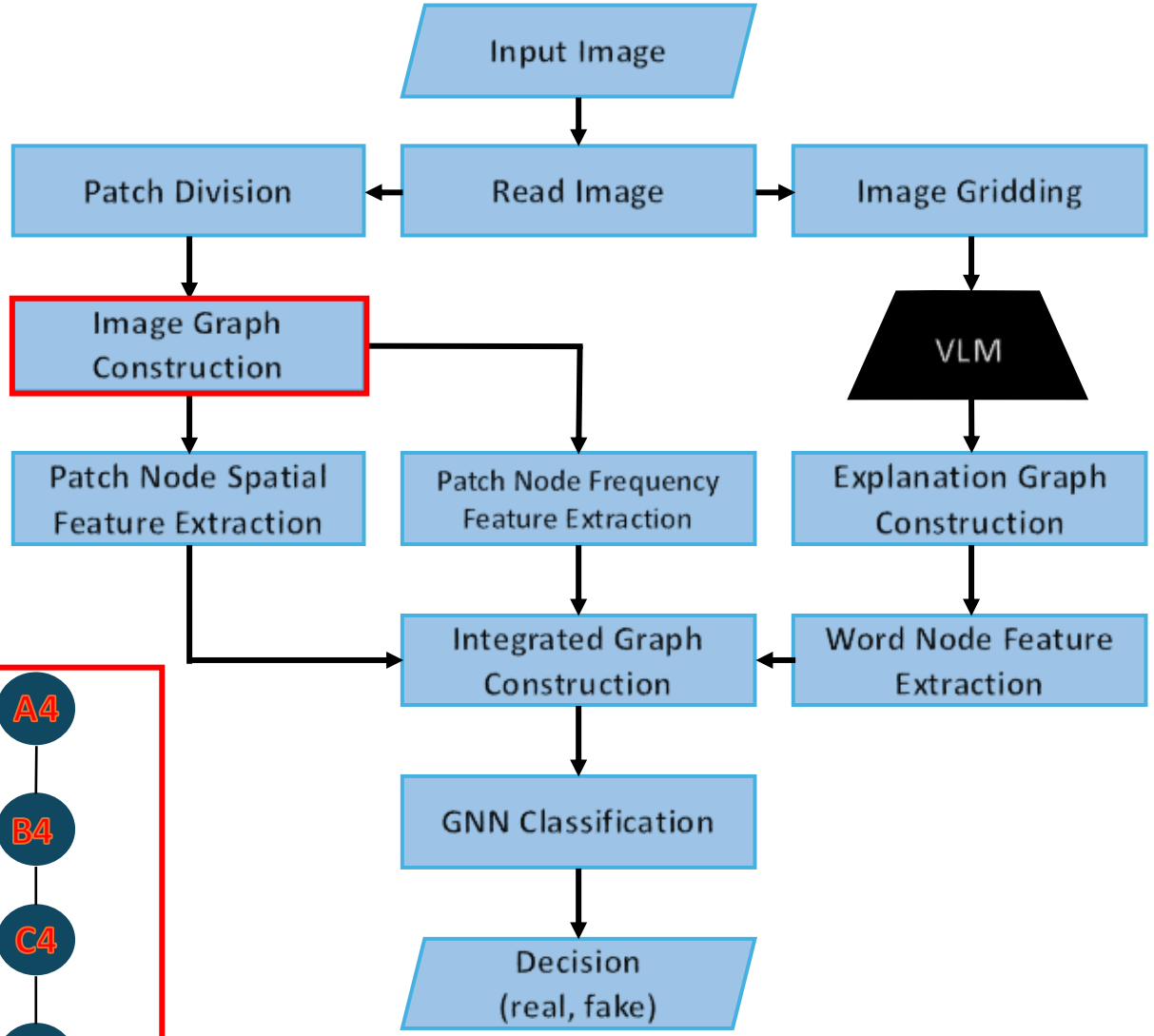




ViGText: The Pipeline

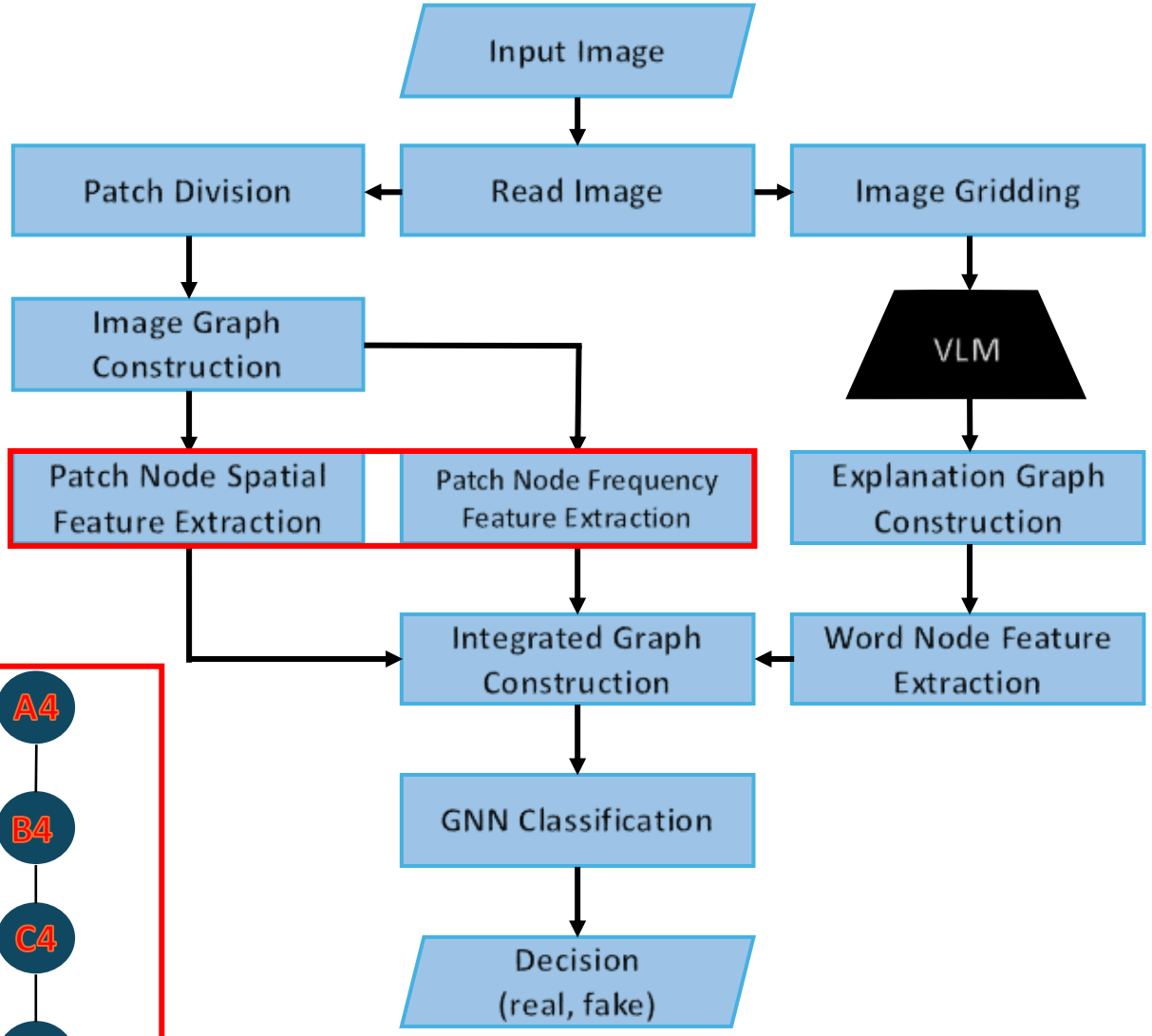
- Initialize an empty image graph
  - Each patch → a graph node
  - Spatial adjacency → graph edges

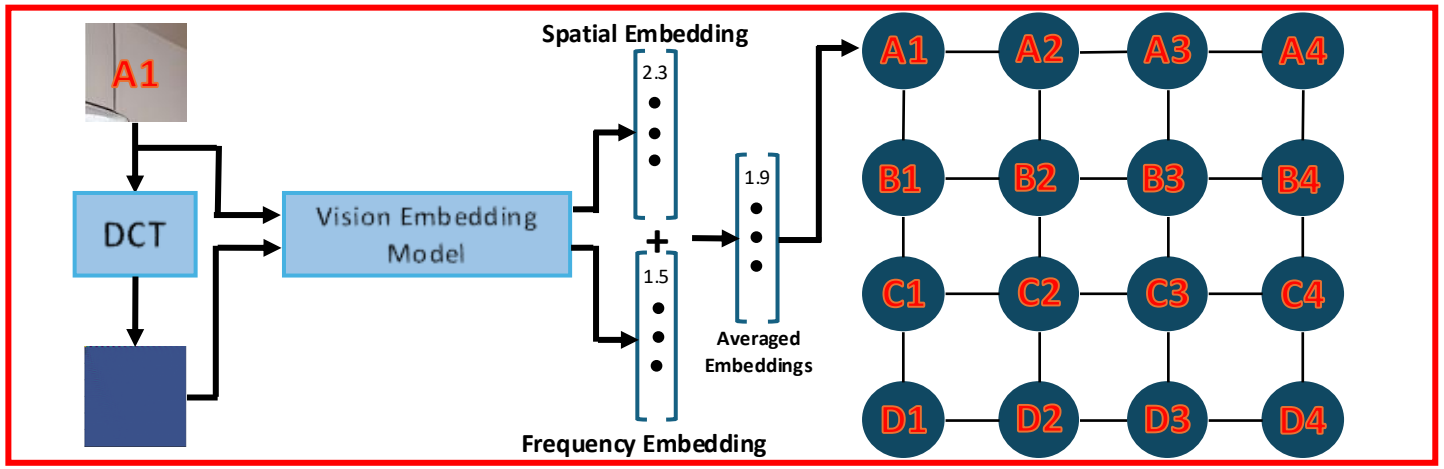




ViGText: The Pipeline

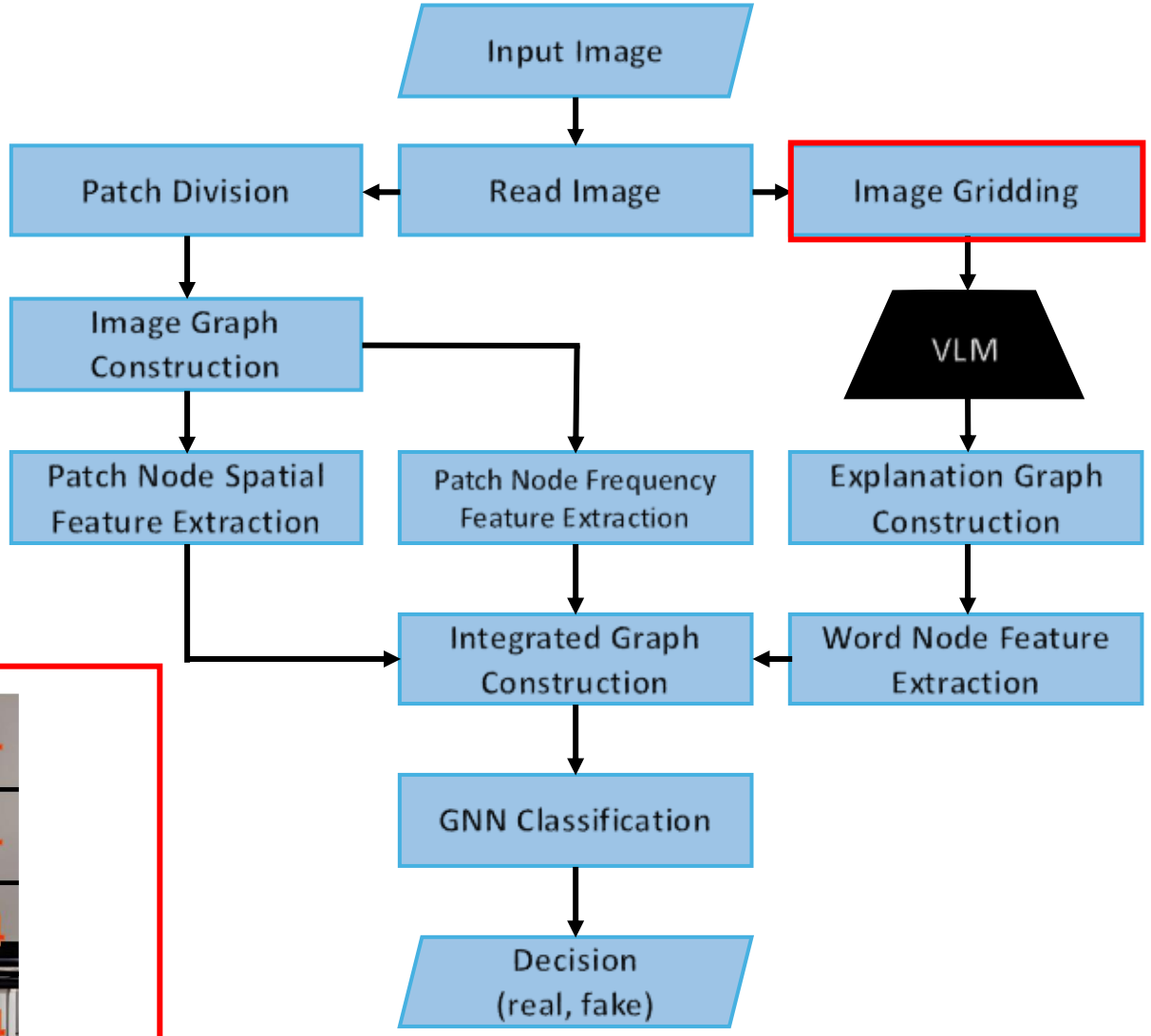
- Patch → DCT transformation
- Spatial + frequency features embedded
- Node representation constructed





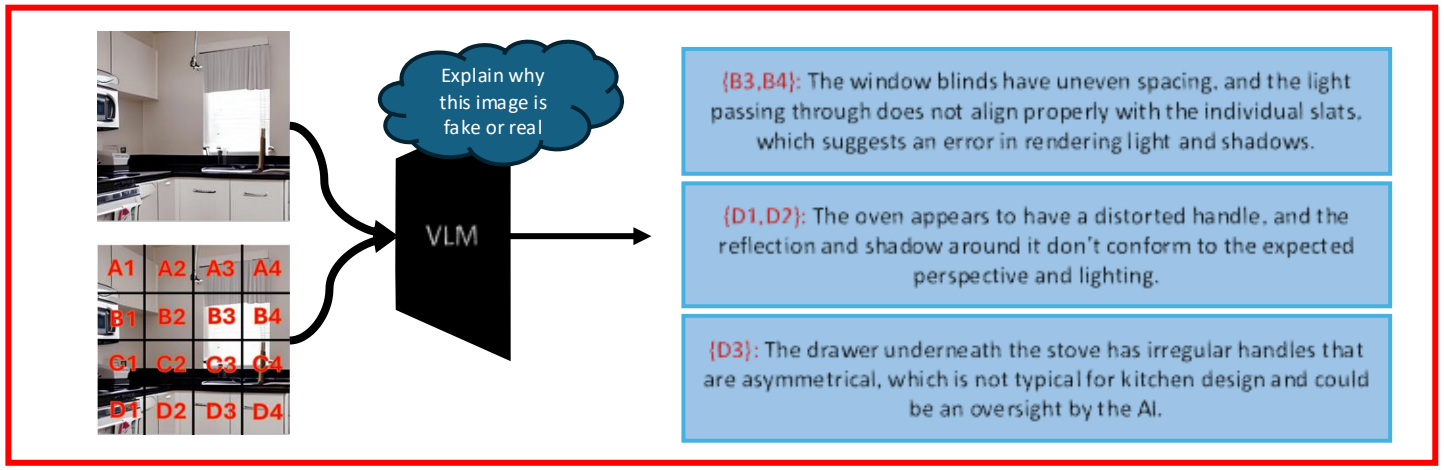
ViGText: The Pipeline

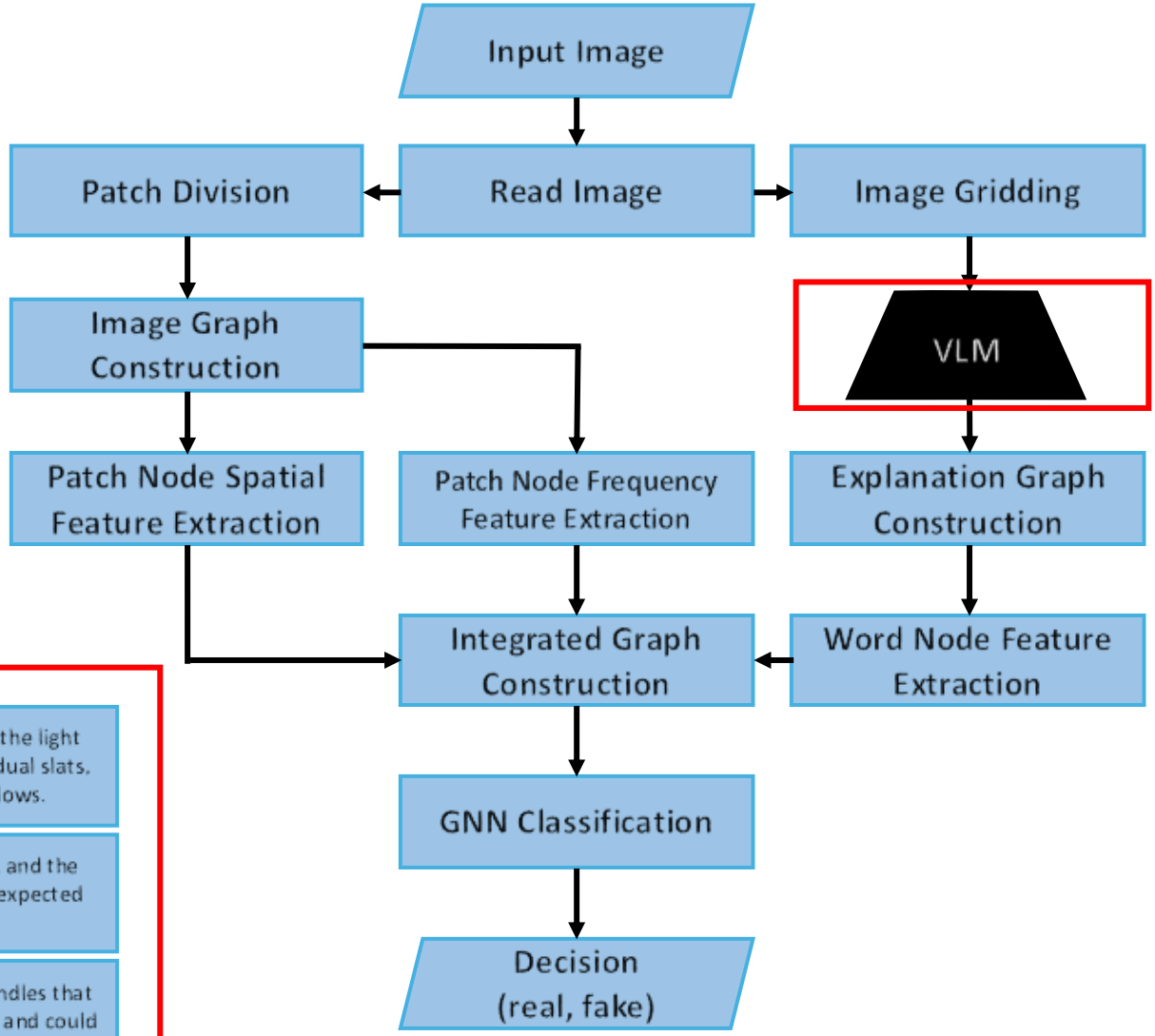
- Grid-based overlaying of the input image
- 16 equal patches for explanation generation



ViGText: The Pipeline

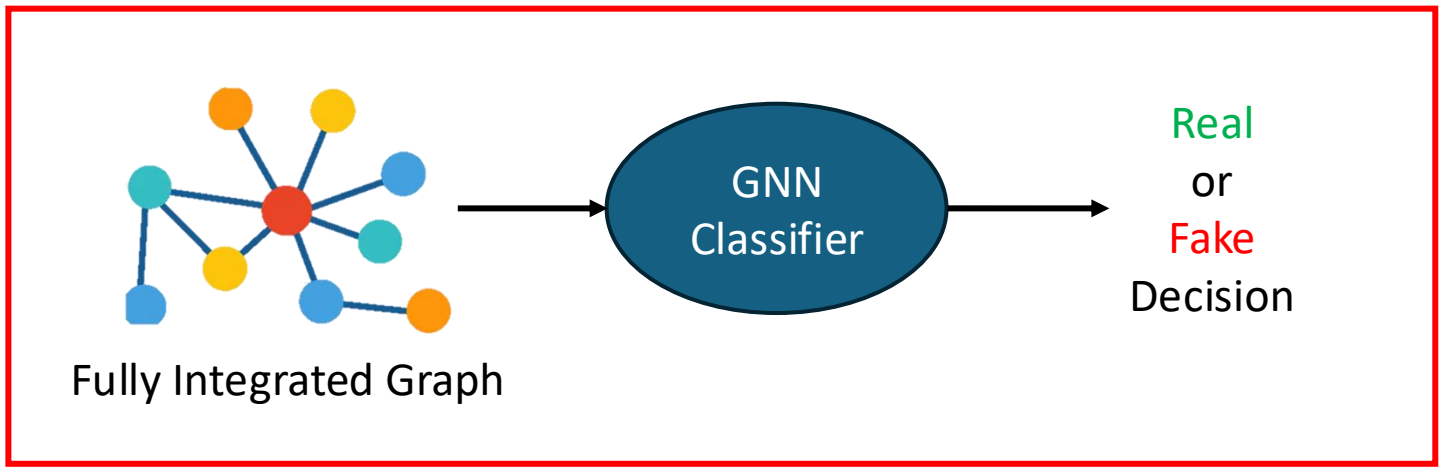
- Input image + grid overlay → VLM
- Generates patch-grounded explanations

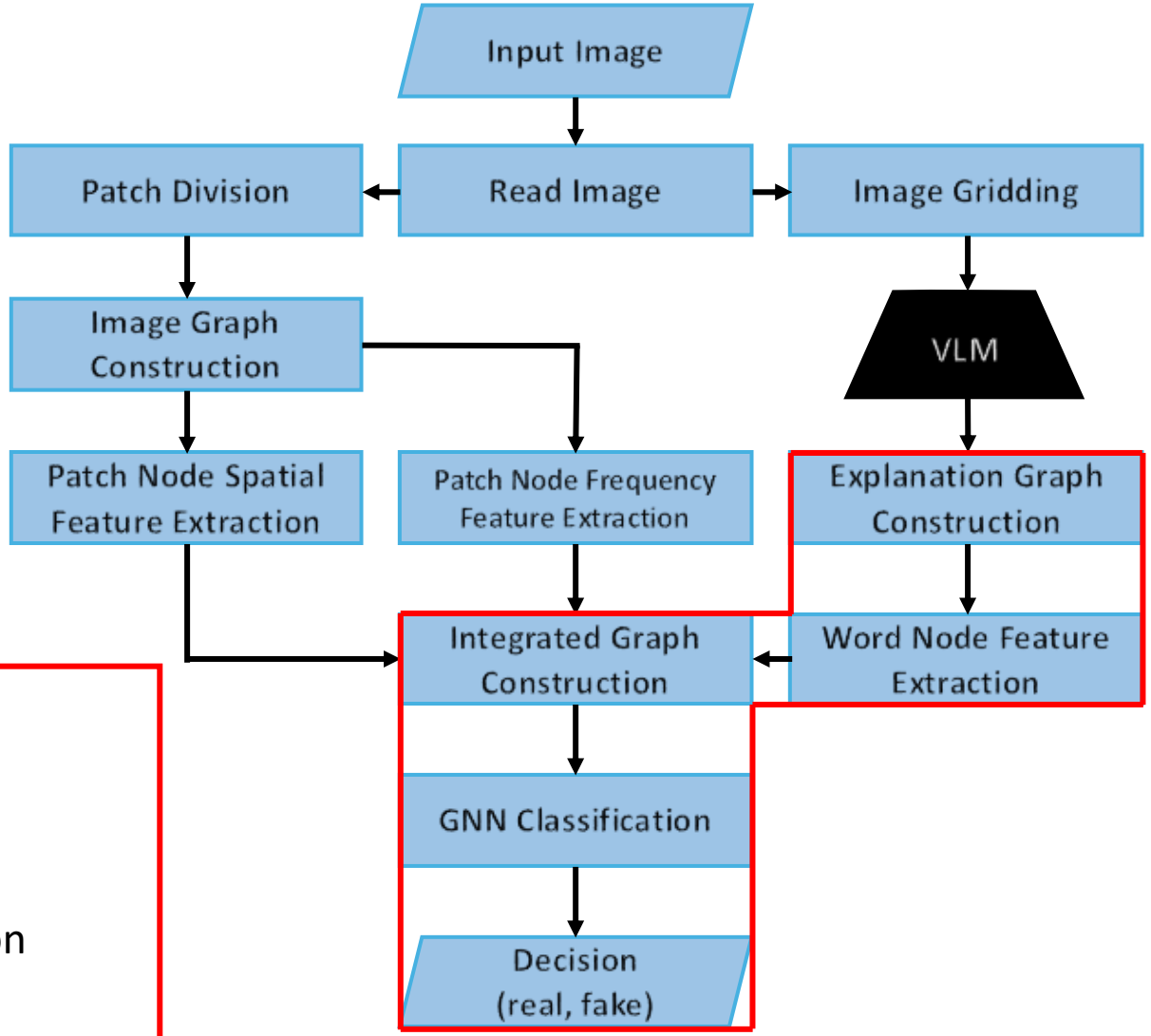




ViGText: The Pipeline

- Build explanation graph
  - Words → nodes
  - Syntax → edges
- Embed word nodes
  - Text embeddings as node features
- Integrate with image graph
  - Connect words ↔ patches
- Classify with GNN
  - Real / Fake





ViGText: Experimental Setup

📁 Datasets:

- **Stable Diffusion (SD):**
  - 8k real (LAION-Aesthetics) + 8k fake (SD1.4) for training.
  - **16 fine-tuned SD1.5 test sets** (8 FM, 8 LoRA) for **generalization** evaluation.
- **StyleCLIP:**
  - 8k real (Flickr-Faces-HQ) + 8k fake (StyleGAN2) for training.
  - **3 adversarially manipulated** test sets for foundation model attack **robustness** evaluation.
- **New SD 3.5 Benchmark:**
  - 8 additional fine-tuned SD3.5 test sets introduced in this study to extend the **generalization** evaluation.

📊 Baselines:

- **DE-FAKE [2]:** CLIP image + text prompt embeddings → MLP.
- **DCT [3]:** Frequency-domain features + logistic regression.
- **UnivCLIP [4]:** Universal CLIP-based detector.

LoRA: Low Rank Adaptation
FM: Full Model, SD: Stable Diffusion

ViGText: General Performance

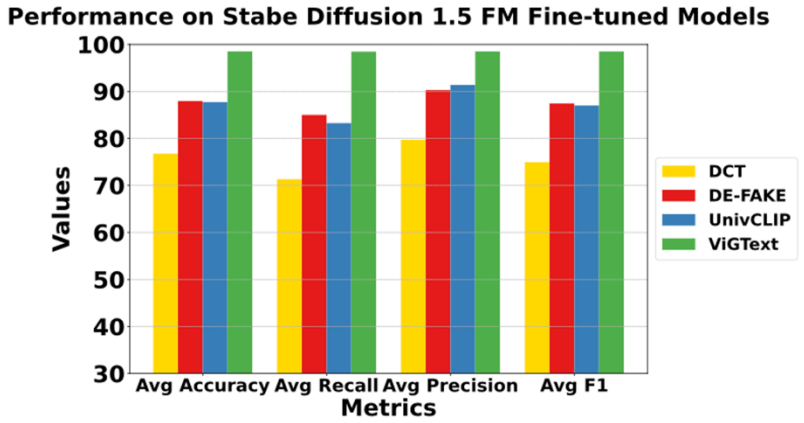
- Strong performance on in-distribution test sets
- Nearly all methods report high classification metrics

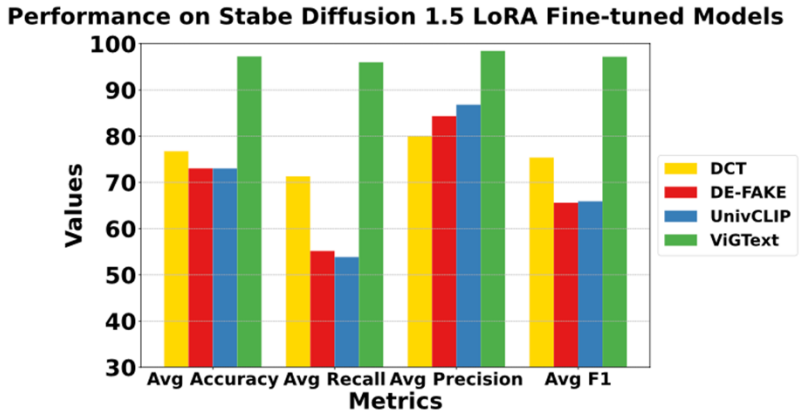
TABLE 3. PERFORMANCE ANALYSIS ON THE RESPECTIVE TESTING SETS OF THE DATASETS (THE HIGHEST IS IN BOLD).

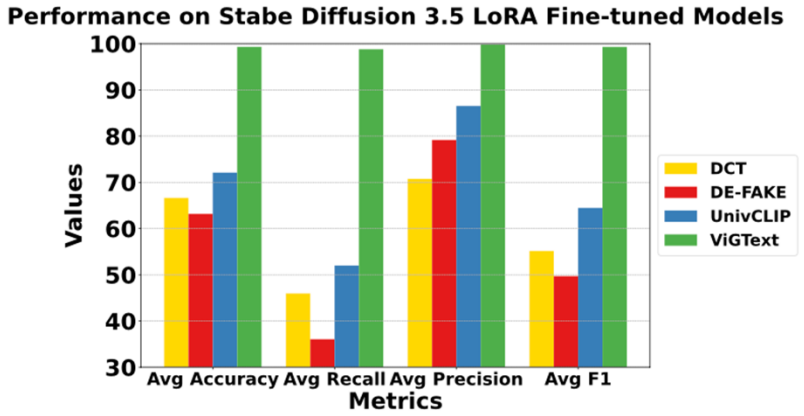
Approach	SD				StyleCLIP			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
DCT	85.50	83.30	88.80	85.96	98.80	98.22	**99.40**	98.80
DE-FAKE	92.45	91.17	94.00	92.5	74.05	75.34	71.50	73.37
UnivCLIP	93.04	92.33	93.89	93.10	93.04	93.79	92.19	92.99
ViGText	**99.25**	**99.8**	**98.52**	**99.26**	**99.60**	**99.90**	99.21	**99.60**

ViGText: Generalization Performance

- **Scenario**
  - **Train**: SD 1.4 fakes
  - **Test**: SD 1.5 & SD 3.5 fine-tuned variants (LoRA, FM)
- **Generalization Performance**
  - ViGText outperforms recent detectors
  - Maintains near-perfect accuracy on unseen deepfakes





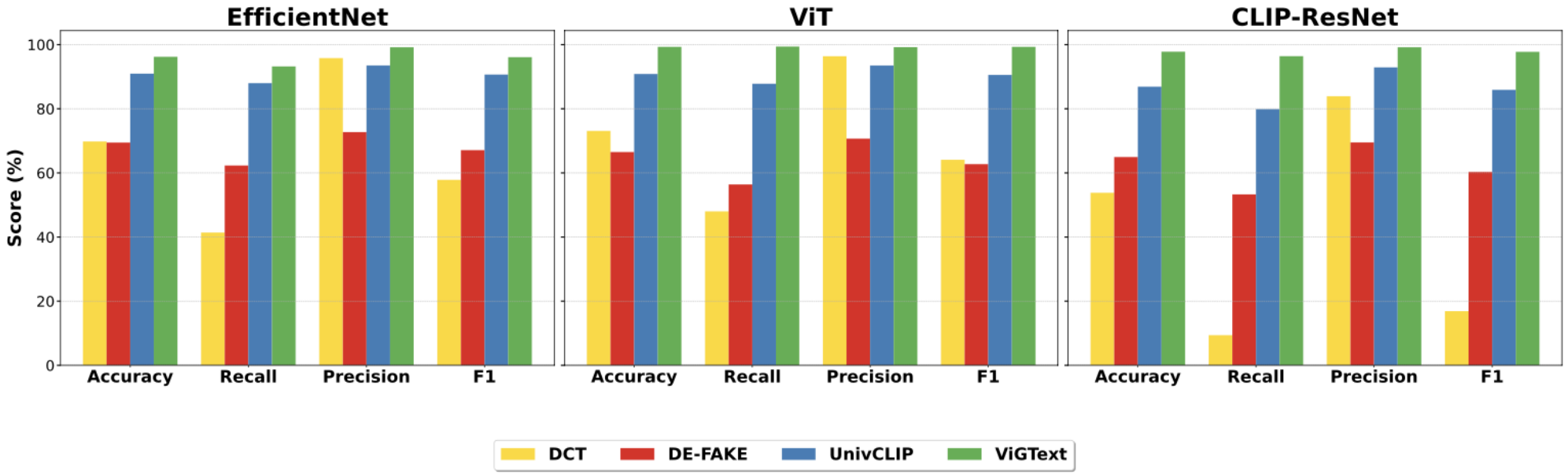


LoRA: Low Rank Adaptation
FM: Full Model, SD: Stable Diffusion

ViGText: Robustness Performance

- **Scenario**
  - **Train:** StyleCLIP fakes
  - **Test:** Adversarial variants

    (Foundation model surrogate)
- **Robustness Results**
  - ViGText outperforms baselines
  - Robust under Foundation model attacks



ViGText: Conclusion and Future Work

- **Conclusion**:
  - Vision–language graph reasoning for deepfake detection
  - Context-aware image–text integration
  - State-of-the-art generalization performance
  - Robust to foundation model attacks
- **Future Work:**
  - Adaptive patching and dynamic graph alignment
  - Extension to video and audio modalities
  - Advanced spectral-domain modeling and spectral LLM reasoning

References

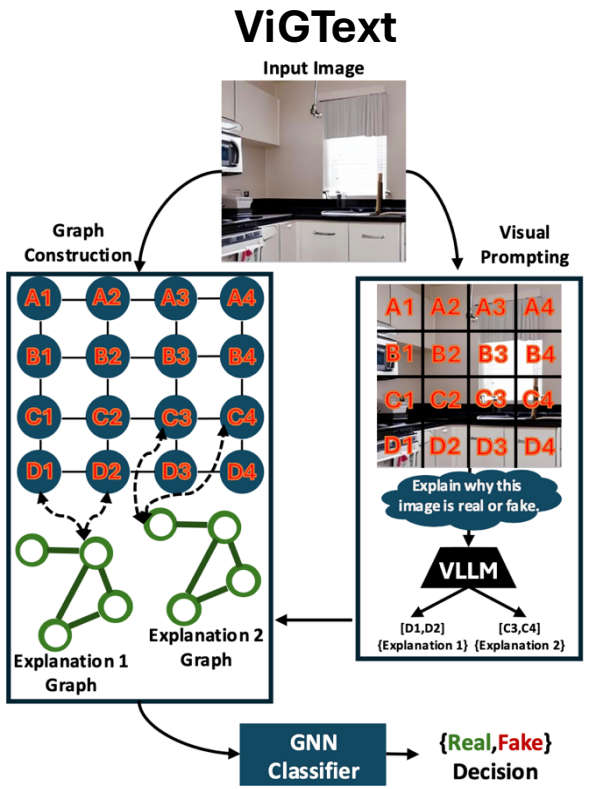
[1] Abdullah, S. M., Cheruvu, A., Kanchi, S., Chung, T., Gao, P., Jadliwala, M., & Viswanath, B. (2024, May). An analysis of recent advances in deepfake image detection in an evolving threat landscape. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 91-109). IEEE.

[2] Sha, Z., Li, Z., Yu, N., & Zhang, Y. (2023, November). De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security* (pp. 3418-3432).

[3] Ricker, J., Damm, S., Holz, T., & Fischer, A. (2022). Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*.

[4] Ojha, U., Li, Y., & Lee, Y. J. (2023). Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 24480-24489).

Contact Information and Resources



- Code and Artifacts:

https://github.com/AhmadALBarqawi/ViGText

- Contact:

Ahmad ALBarqawi

New Jersey Institute of Technology

aka87@njit.edu

NEW JERSEY INSTITUTE OF TECHNOLOGY
Founded 1881