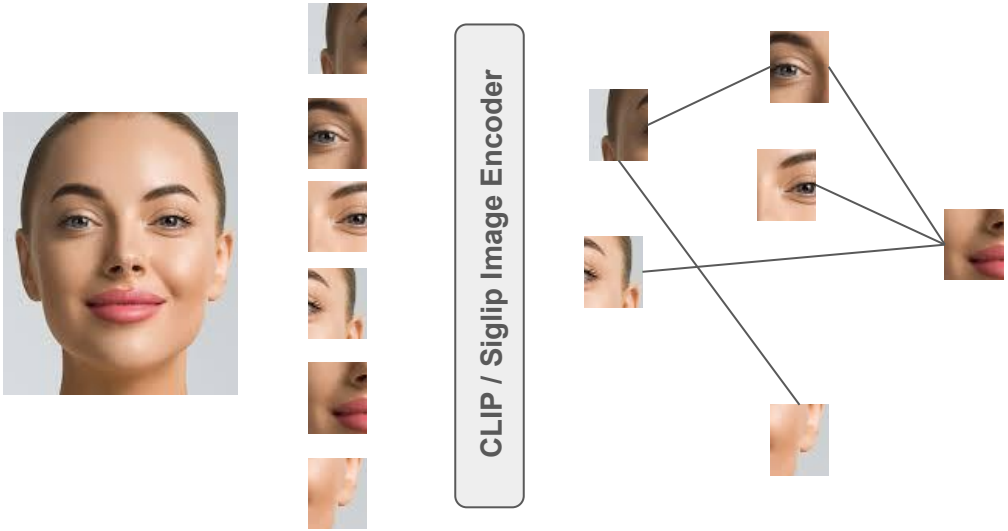
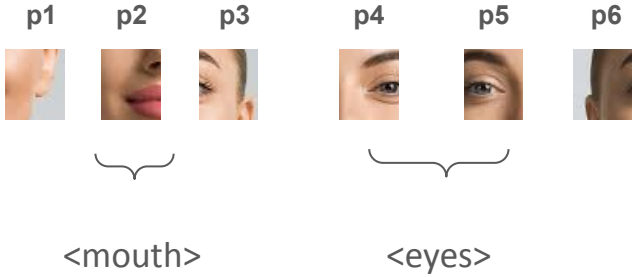
The irregularity of facial region shapes presents a significant challenge for conventional techniques, such as grid-based CNN architectures or sequence-based Transformer models, when processing facial images. [1] advocates for the use of graph representation in deepfake detection, given its ability to capture the diverse topologies and geometries of facial regions.

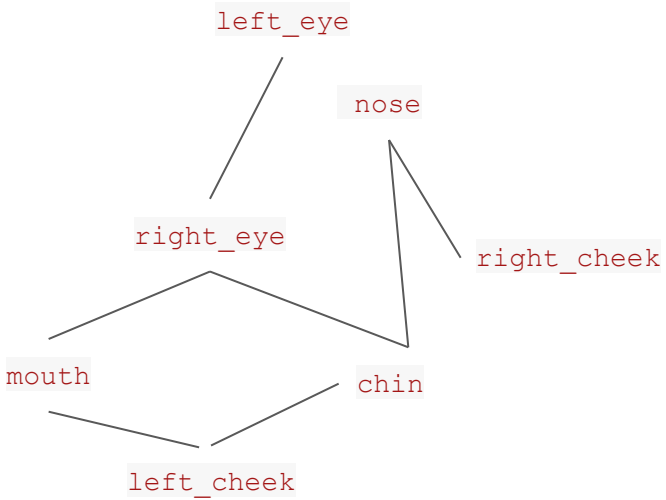


The nodes are feature embeddings extracted from the image patches.
The adjacency relationship between patches forms the graph edges, which are established based on a spatial proximity criterion that connects each patch with its neighboring patches.

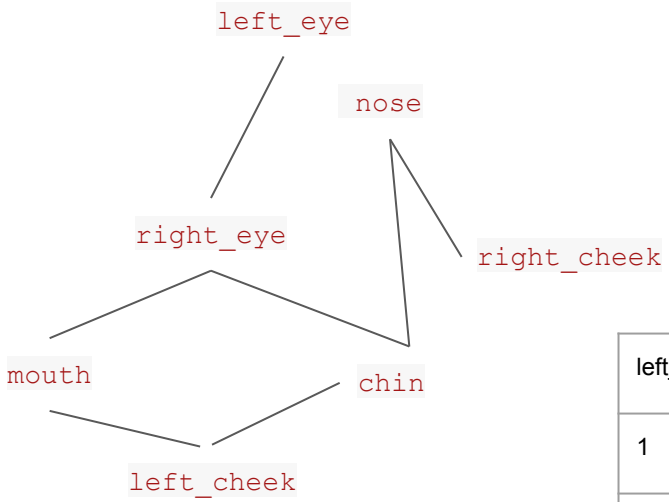
[1] Khormali, Aminollah, and Jiann-Shiun Yuan. "Self-supervised graph Transformer for deepfake detection." *IEEE Access* (2024).

Utilizing landmark coordinates from the preprocessing phase allows us to pinpoint patch locations for each facial parts.[2]





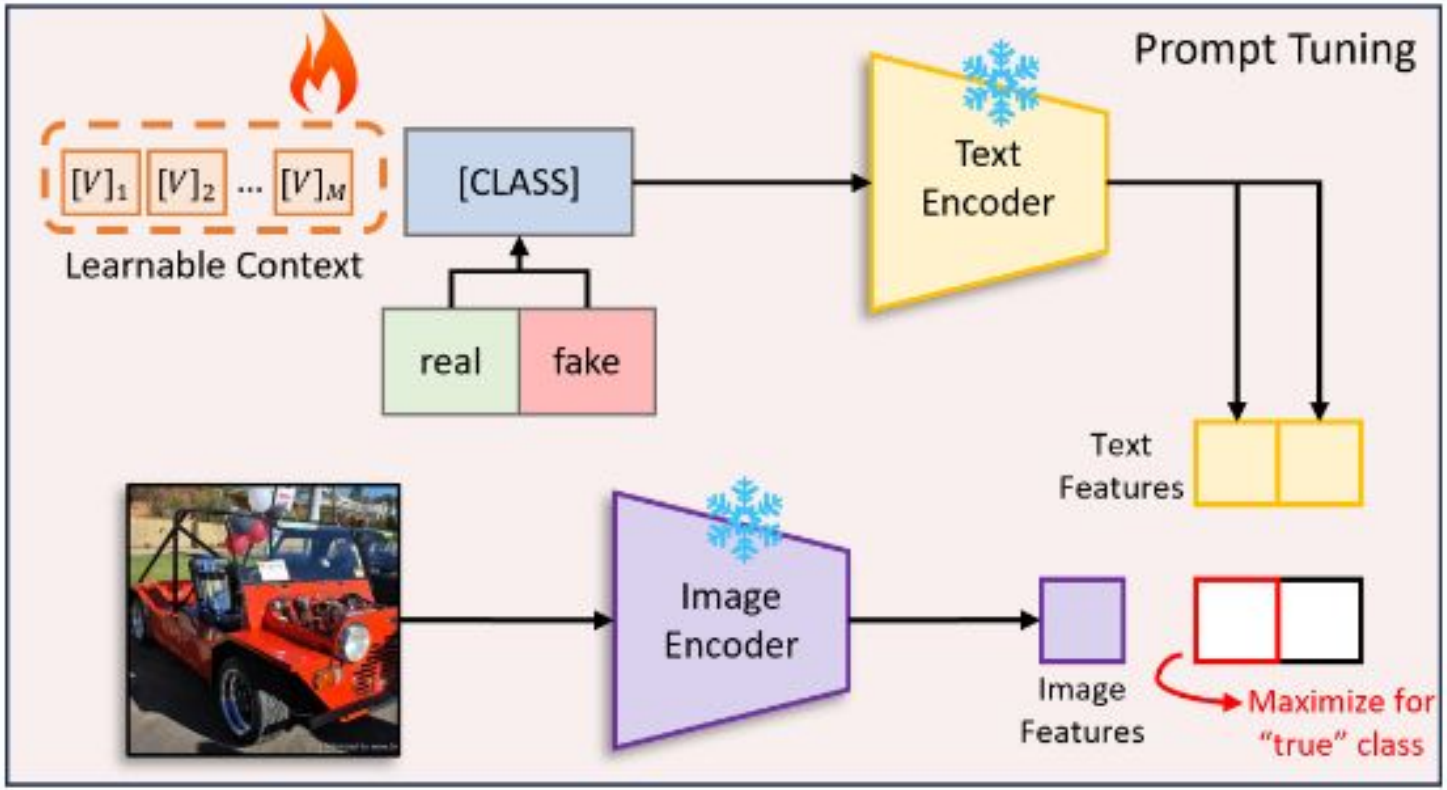
[2] Han, Yue-Hua, et al. "Towards More General Video-based Deepfake Detection through Facial Feature Guided Adaptation for Foundation Model." *arXiv preprint arXiv:2404.05583* (2024).



left_eye	right_eye	nose	right_cheek	chin	Left cheek	mouth
1	5	2	3	0	0	4
0	4	0	5	1	3	2
3	1	0	0	4	2	5

left_eye X right_eye X X mouth X X left_cheek X chin X X X nose X X X X right_cheek

Prompt template



Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. 2024. CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection. In Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24). Association for Computing Machinery, New York, NY, USA, 1006–1015.

Contrastive Vision Encoder

Linear projection

SigLIP: 400M Vision Model

"Where is the photographer resting?"

Tokenizer (SentencePiece)

...

...

Gemma:
2B Language Model

Transformer Decoder

"In a hammock under a tree on a tropical beach"