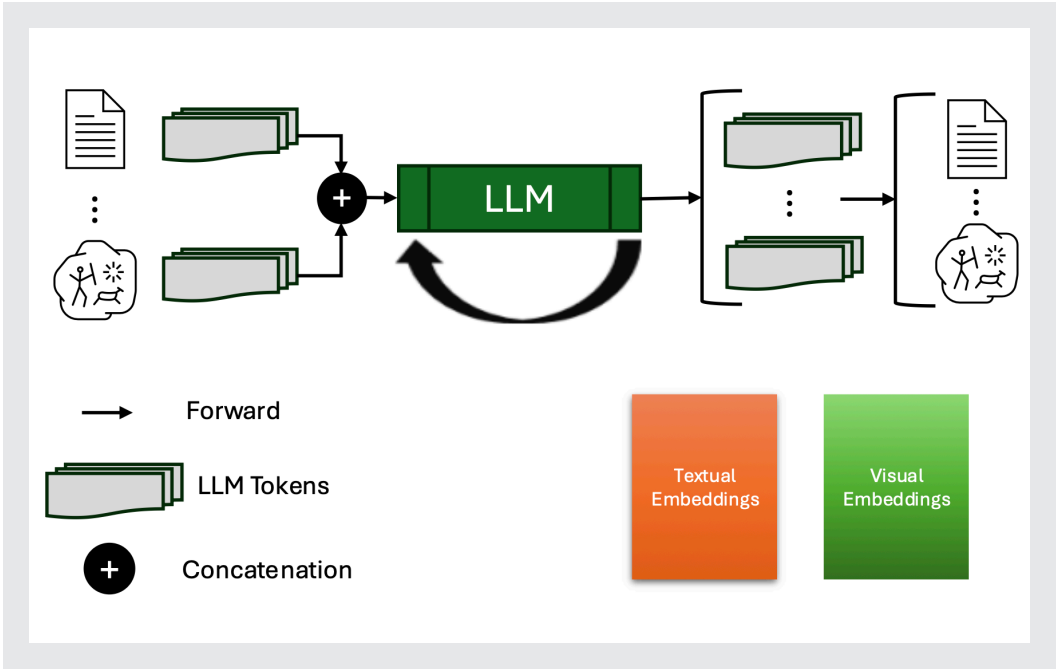
School of Engineering

Institute of Embedded Systems

Multimodal Autoregressive U-Net Transformer Architecture for Pixel-wise Image and Text Conditioning



- Unified Multimodal Architecture: Single transformer processes text and visual tokens autoregressively
- Token-Centric Framework: Visual and textual tokens interleaved and processed identically
- CLIP-Conditioned U-Net: Custom U-Net performs multi-scale visual encoding and decoding
- Flexible Visual Tokenization: Two quantization approaches balance fidelity and computational efficiency
- Multimodal Task Demonstration: Model generates text-conditioned segmentations with zero-shot capabilities

Master Thesis
Author: Dávid Isztl
Supervisor: Prof. Dr. Matthias Rosenthal
www.zhaw.ch/ines