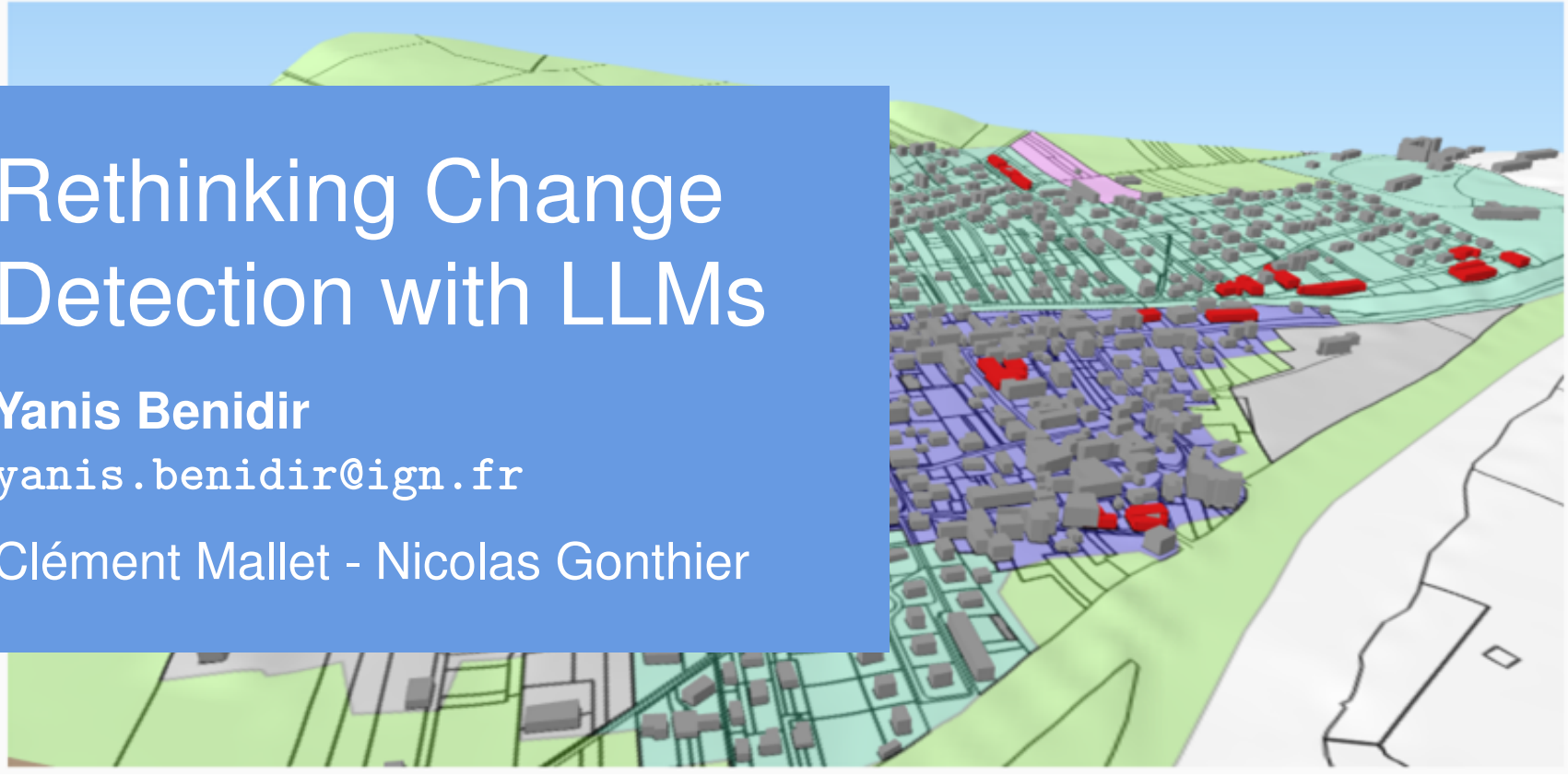
Rethinking Change Detection with LLMs

Yanis Benidir
yanis.benidir@ign.fr

Clément Mallet - Nicolas Gonthier

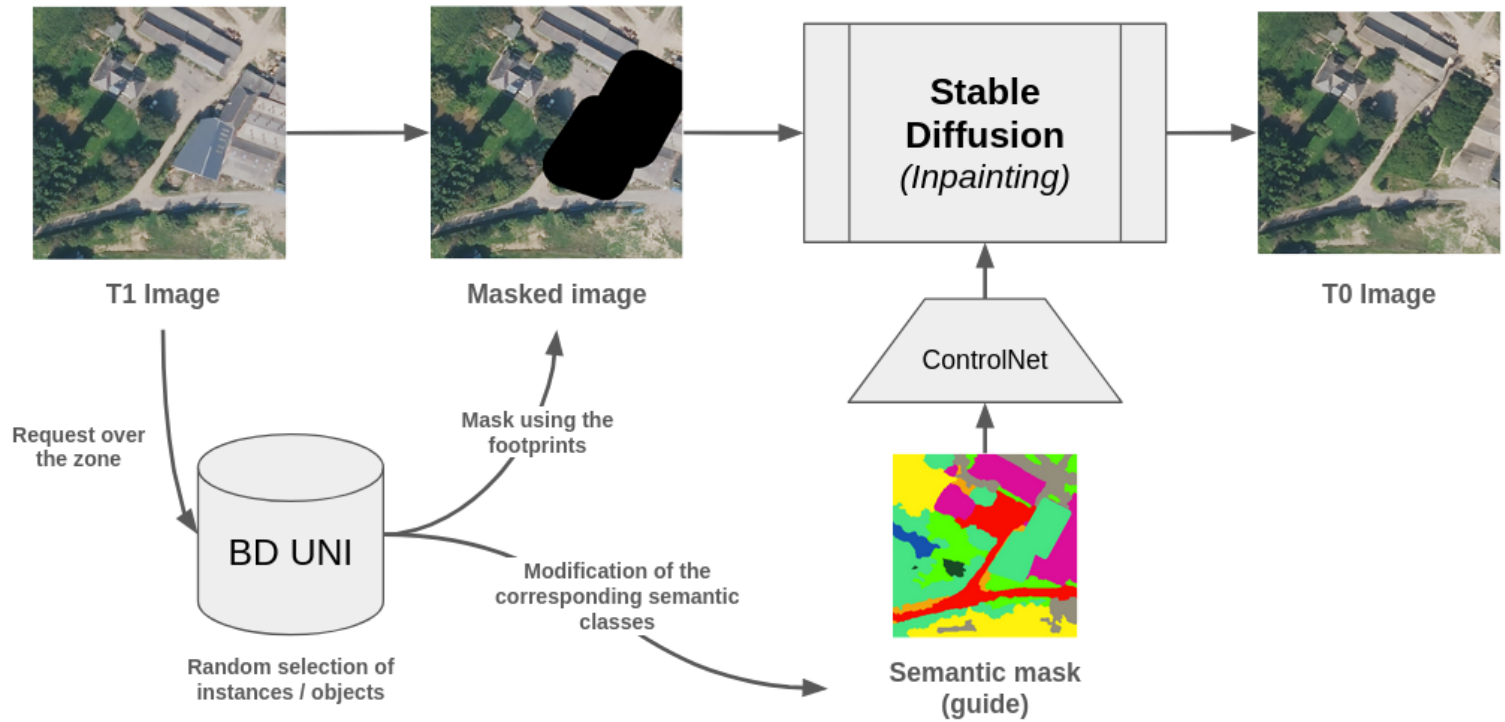


Outline

Outline

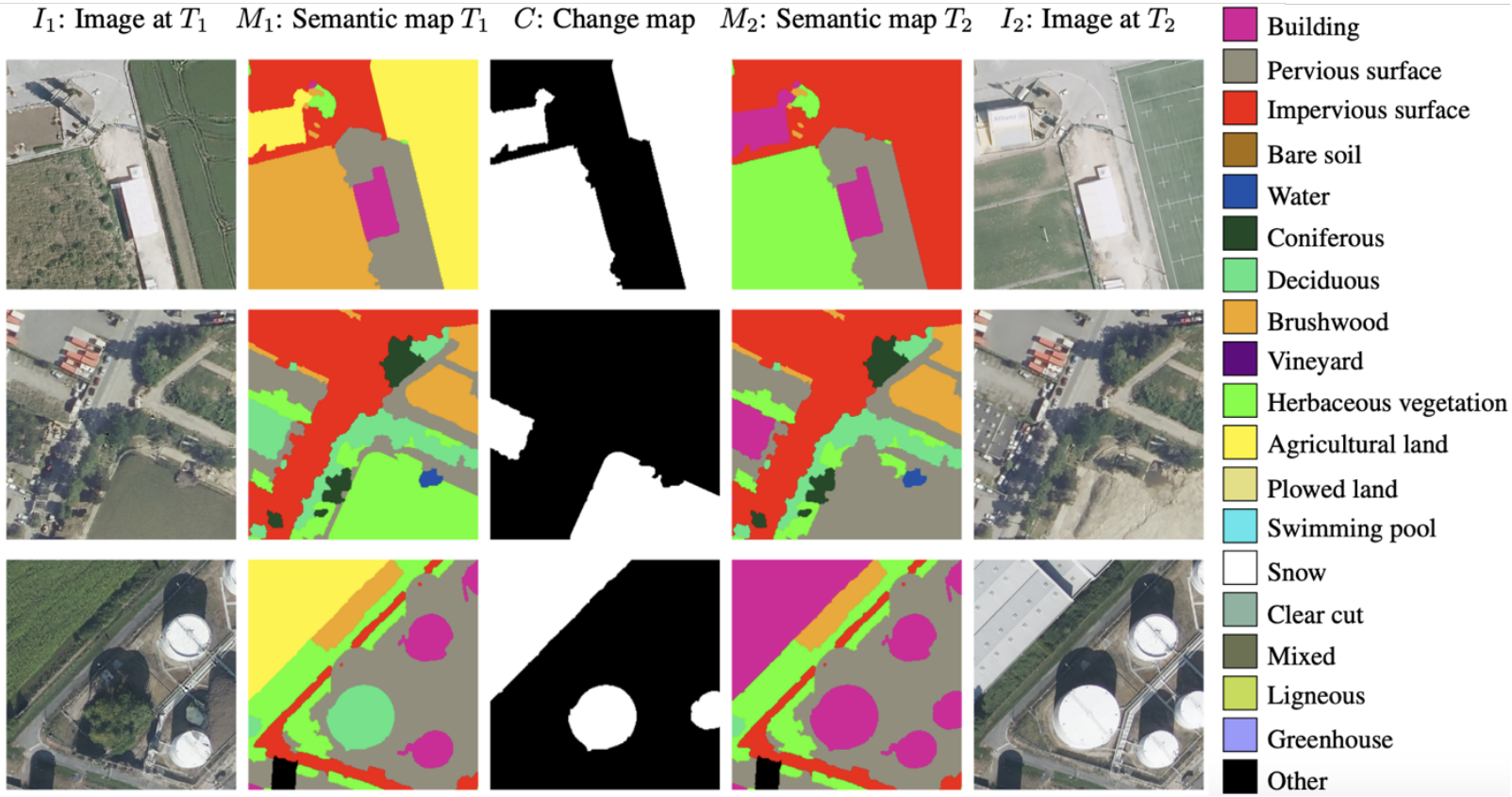
Our generation pipeline : HySCDG

"*top view of coniferous and deciduous next to a road in the locality of Vendresse, Grand Est, in the morning, during summer, high resolution, highly detailed*"



HySCDG, our generation pipeline - Process "*Select, mask, change, inpaint*"

FSC-180k Dataset



FSC-180k consists of 180k semantically-annotated images pairs

---

Yanis Benidir, Nicolas Gonthier, Clément Mallet. **The Change You Want To Detect: Semantic Change Detection In Earth Observation With Hybrid Data Generation**, CVPR 2025

Outline

Rethinking Change Detection

Aligning Model Capabilities with Real-World Needs

- Identify strategic use cases (national and regional mapping agencies, ...)
- Implement a **user-in-the-loop** feedback system
- Control and customize model behavior via *prompts*
- Key shifts:
  - (a) Discretization: From "Pixel-level" → "**Instance-level**" reasoning
  - (b) Richer outputs: *Change captioning*

Towards Interactive Models

- **Query-driven** detection (guided by the user's question)
- Dynamic adaptation to the specific task
- Focus on **explaining** the change, not just localizing it

Target Use Cases

Concrete Examples

- Topographic Database updates
- Monitoring urbanization and soil sealing (artificialization)
- Post-disaster damage assessment
- AI-Assisted interpretation of aerial imagery

The Core Need

A single "Universal", Multi-task, Language-Driven Model

The Solution: Vision Large Language Models (VLLMs)

Key Capabilities

- Visual Question Answering (VQA)
- **Multimodal** reasoning (Images + Text)
- Deep semantic understanding
- Natural Language Interaction via **prompting**

Main Objective

Answer "Change-related" questions based on image pairs/sequences

A Multi-Task Approach

Functionalities

- Classic VQA Mode: Textual Questions / Answers
- Advanced Features:
  - Adapted specifically for Remote Sensing
  - **Image Generation:** Semantic maps / Change maps
  - **Instantiation:** Objects detection and segmentation
  - **Spatial Analysis:** Counting, Measuring areas

The "Universal" Dimension

Open Vocabulary Change Detection (OVCD)

- Unspecified classes, extensible vocabulary
- Classes represented by **text embeddings** rather than fixed labels

Outline

Global Architecture

Backbone (Starting Point)

- *Vision Encoder*: SigLIP[1]
- LLM: LLaMA / Qwen2[2]
- Base Architecture: **VideoLLaMA3**[3]

Proposed Improvements

- Fine-tuning specifically for Remote Sensing + Change Detection
- Integration of a custom **Image Decoder**
- True Multimodal Output: Text + Images

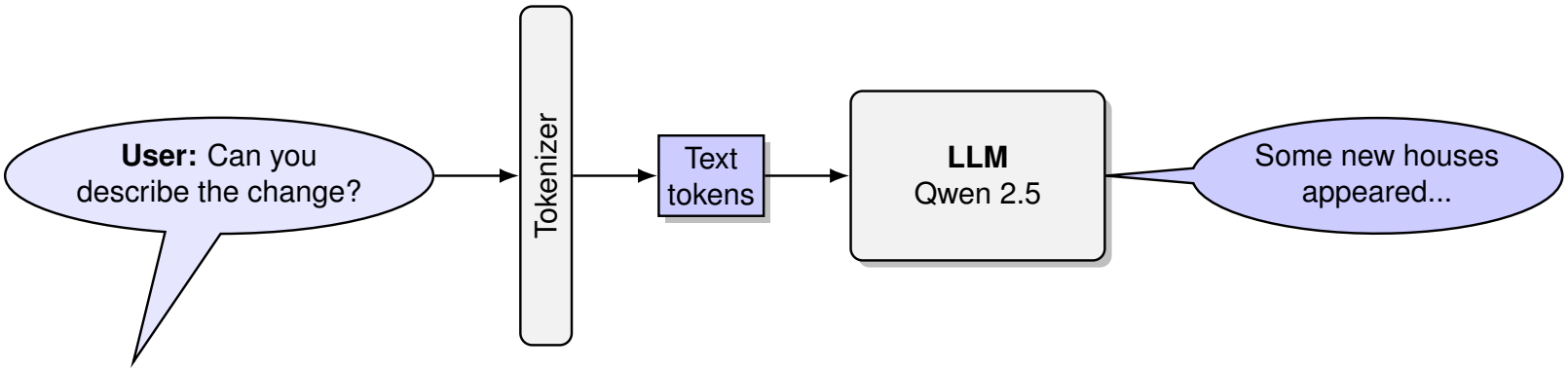
[1] Zhai et al., *Sigmoid Loss for Language Image Pre-Training* (ICCV 2023).
[2] Yang et al., *Qwen2 Technical Report* (2024)
[3] Cheng et al., *VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs* (2024).

Our multimodal architecture

User: Can you describe the change?

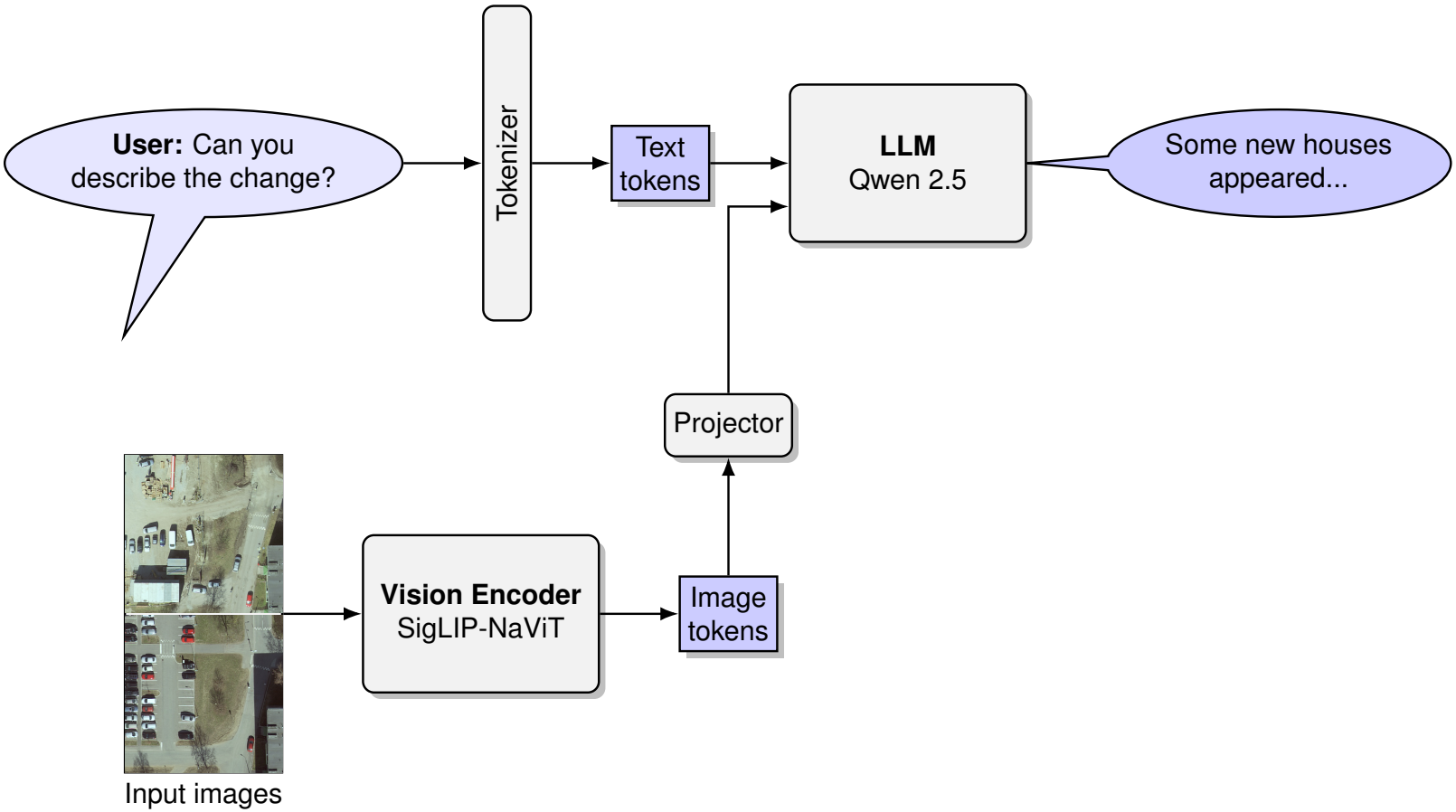
Input images

Our multimodal architecture

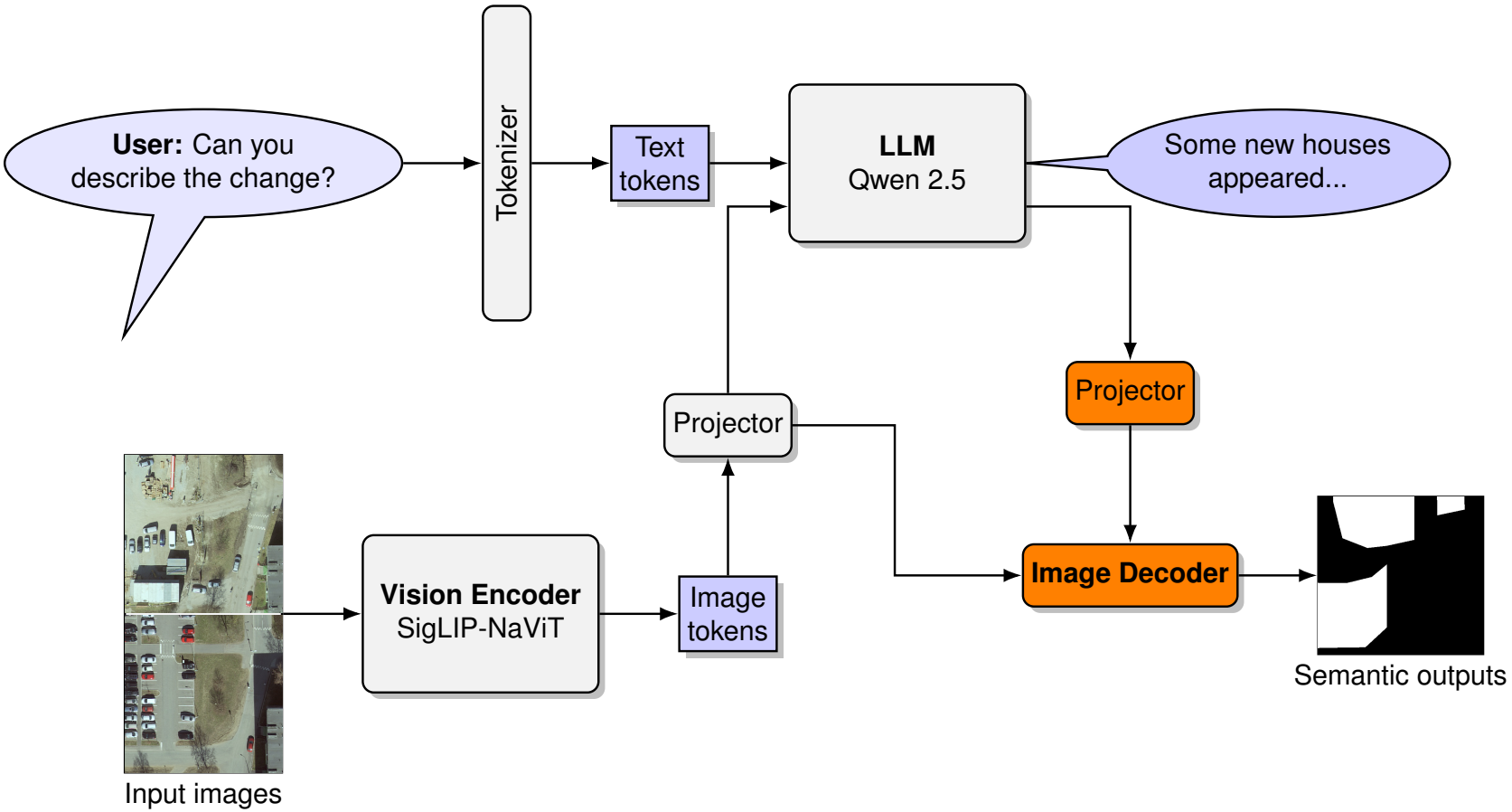


Input images

Our multimodal architecture



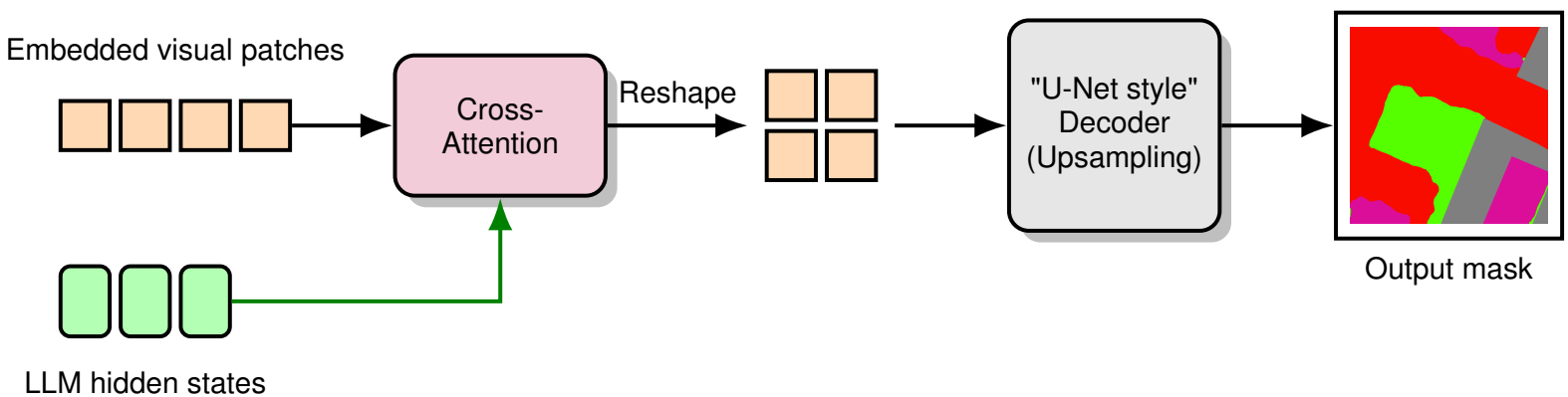
Our multimodal architecture



The Image Decoder

Mechanism

1. **Fusion:** LLM features integration → Cross-Attention on linear sequences
2. **Reshape:** Linear → 2D Mini-Image
3. **Decoding:** Upsampling / U-Net blocks



Training vs. Inference Alignment

The Challenge

- **Training:** Access to the full ground truth sequence
- **Inference:** Token-by-token autoregressive generation

The Solution

- A special learnable token: `<image_latent>`
- The token's *hidden state* captures the necessary semantic info
- The model learns to predict this token
- Appearance of the token triggers the Image Decoder

Outline

Training Strategy

Loss Functions

- Text Loss: *Cross-Entropy*
- Semantic Loss: *Cross-Entropy*
- Change Loss: *Cross-Entropy* + *Dice Loss*
- **Weighted** losses (imbalanced classes)

Training Methods

- Modules are at different pre-training stages
- Sequential vs. Simultaneous training (+ LoRA ?)
- Specific learning rate scheduling per module
- Challenge: **High numerical instability**

Training Data

Building a Multimodal Dataset

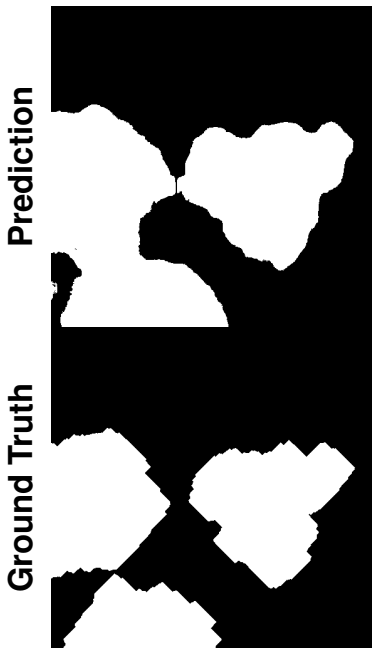
- Combining "Classic" Change Detection data + Captioning
- Aggregating diverse datasets:
  - FSC-180k, HiUCD, SECOND-CC
  - xView2 (Natural Disasters)
  - LEVIR-MCI

Question Types Generation

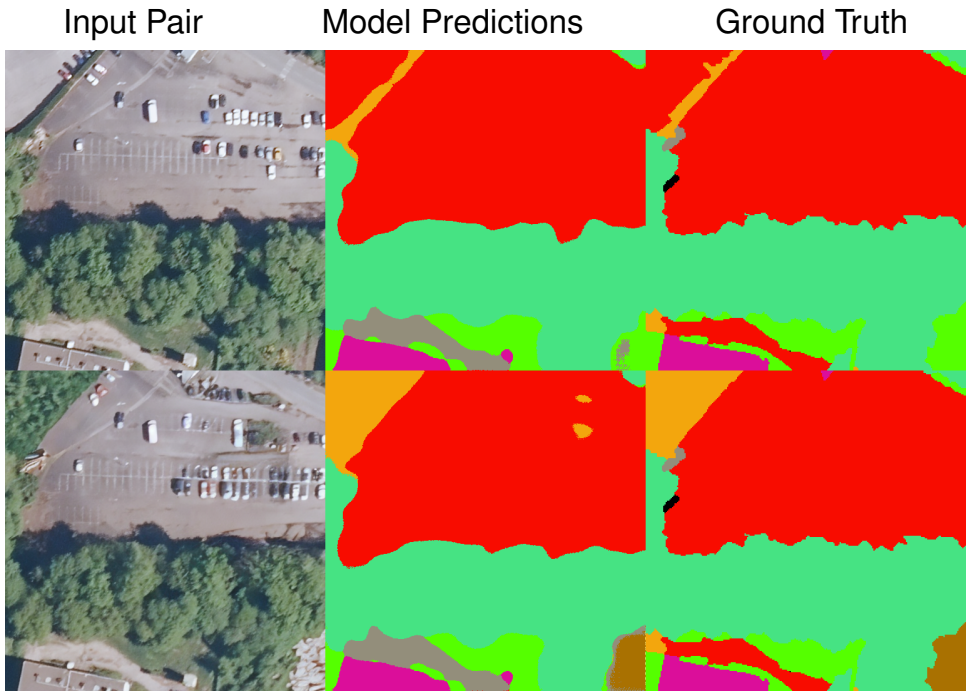
- Specifying semantic classes
- Identifying change trajectories (What happened?)
- Counting / Instantiation tasks
- In Practice: Automated pipeline to generate questions from existing annotations

Examples and Results

Binary Change



Semantic segmentation

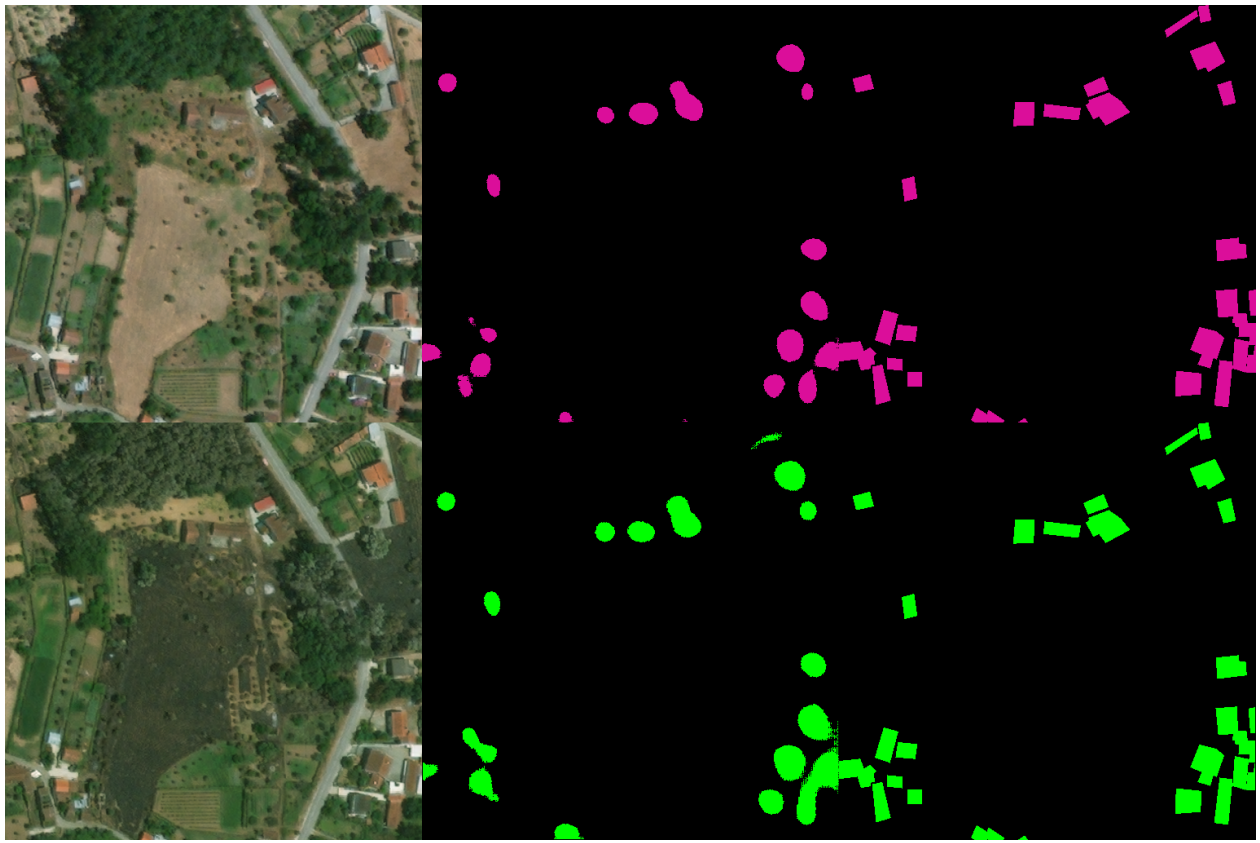


Q: What has the road replaced?
A: The new road replaced the deciduous trees.

Q: How has evolved this zone?
A: Nothing has changed. There are only some temporary changes.

Q: What is the area of the change?
A: The total area of the change is 1750sqm. The deciduous trees was replaced by agricultural vegetation on a surface of 429sqm. The grass was replaced by water on a surface of 836sqm. The grass was replaced by brushwood on a surface of 484sqm.

Thank You



yanis.benidir@ign.fr