Dynamic Sub-Region Search in Homogeneous Collections Using CLIP

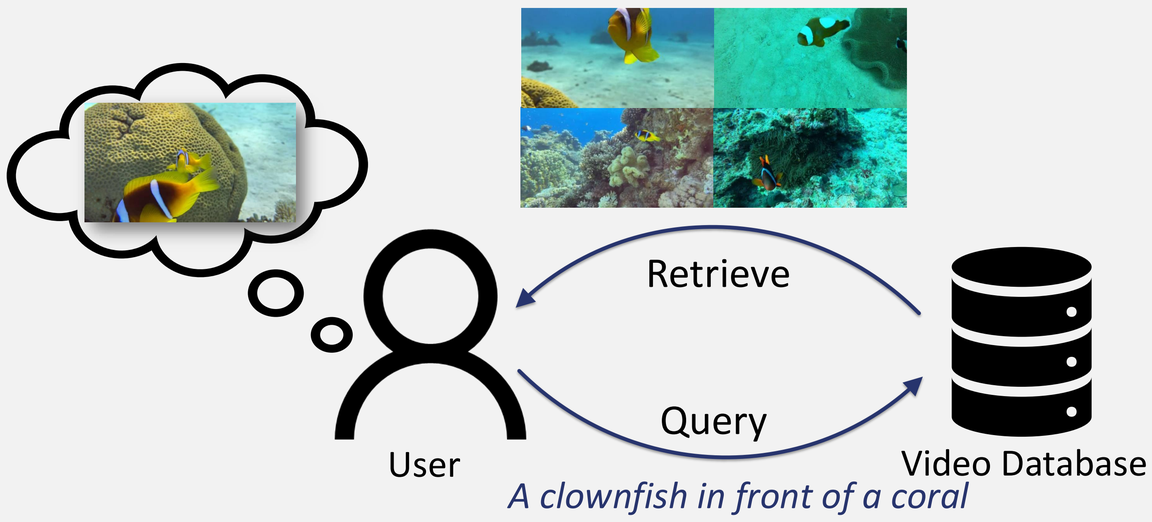

Bastian Jäckl
bastian.jaeckl@uni-konstanz.de


Vojtěch Kloda
vojtech.kloda825@student.cuni.cz


Daniel A. Keim
keim@uni-konstanz.de


Jakub Lokoč
jakub.lokoc@matfyz.cuni.cz

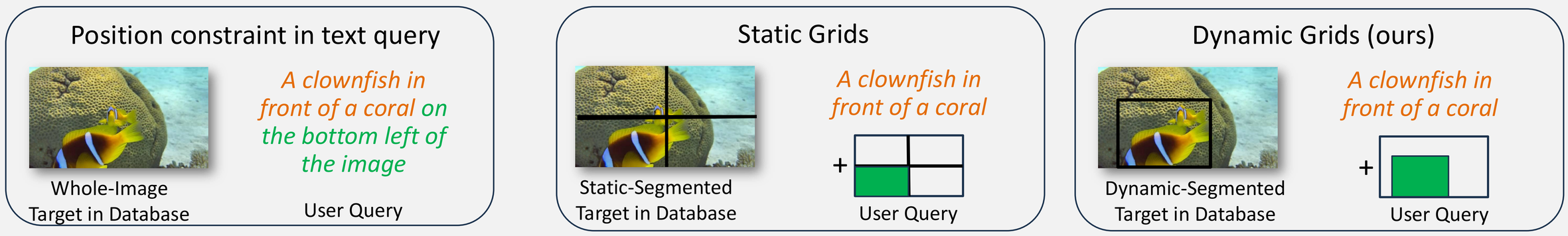
1) Problem: Querying in Homogeneous Collections by Text



- **Multimodal retrieval:** Modern systems embed text and images in a joint space and perform k-nearest neighbors search to retrieve candidates [1].
- **Homogeneous-domain failure:** In homogeneous settings (surveillance videos, medical imagery, underwater scenery) users issue generic queries that match semantically similar but false candidates.
- **Spatial Cues Matter:** Users often know **where** an entity appears, but encoding location in text queries alone yields unsatisfactory matches [2]. We investigate how to explicitly incorporate **spatial constraints into natural-language queries.**

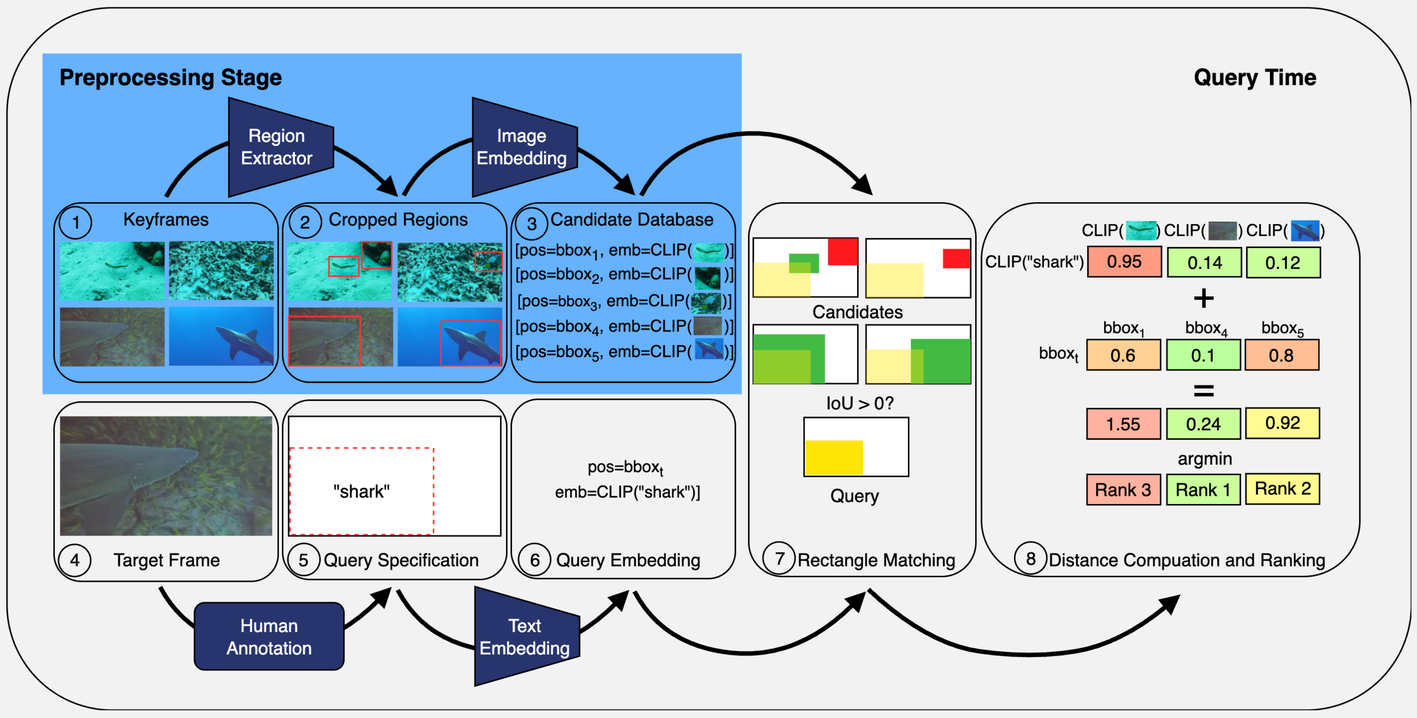
2) Proposed Approach: Dynamic Sub-Region Search

How to incorporate **spatial constraints** **into** **semantic text queries**?



Dynamic Grids: Approach & Implementation

1. Starting Point: Database of **non-annotated images** (or video keyframes.
2. Automatically **segment** each image into **semantically coherent** candidate **regions** using open-set models (e.g., WaterMask, SAM2, Grounded-SAM2).
3. For every region, compute and store its **CLIP embedding** together with its bounding box.
4. At query time, aim to retrieve the most relevant database item for the user's target.
5. The user provides a **text description** plus a **query bounding box** indicating where the description should match
6. Encode the text into a **CLIP text embedding**.
7. Pre-**filter candidates** by selecting all indexed regions whose boxes **overlap the query box**.
8. Rank remaining candidates by **fusing semantic distance** (with CLIP) and **geometric distance** (e.g., IoU/centroid/shape/area).



3) Evaluation and Results

- **Setup & metrics:** Evaluated on 84,309 underwater keyframes [3] with **741 human-annotated text queries + bounding boxes** [4].
- **Dynamic vs. Static Grid**: Using dynamic region proposals **without geometry does not outperform static grids** despite higher IoU with annotations. Tight alignment alone is insufficient.
- **Semantics + geometry wins:** Fusing **CLIP cosine distance with IoU distance** delivers the largest retrieval gains: **doubling recall** over whole-image/static baseline.
- **Robustness**: Under box perturbations such as area changes, performance drops but rectangle-based distances **remain above all baselines.**
- **Upper-bound estimate:** Results use **perfect** query boxes as annotators could directly draw boxes into query images. Consequently, they show an upper-bound potential rather than deployed performance.

4) Conclusion

- **Take-Away Message:** Extracting **coherent regions** and fusing **CLIP semantics** with **spatial alignment** (e.g., with IoU/centroid distance) reliably improves retrieval effectiveness in homogeneous domains.
- **Limitation:** Results are upper bounds as we rely on **perfect query boxes:** performance is sensitive to proposal quality and box noise.
- **Future Work:** Broaden validation **across domains**, evaluate in **more realistic user settings** where users only recall scenes from memory, and extend to **video settings** with moving cameras (temporal consistency + region tracking).

[1] Radford et al.: Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning, 2021.
[2] Ranasinghe et al.: Learning to Localize Objects Improves Spatial Reasoning in Visual-LLMVS. Computer Vision and Pattern Recognition, 2021.
[3] Truong et al.: Marine Video Kit: A New Marine Video Dataset for Content-Based Analysis and Retrieval. International Conference on Multimedia Modeling, 2023.
[4] Jäckl et al.: Experimental Evaluation of Static Image Sub-Region Based Search Models using CLIP. Similarity Search and Applications, 2025.

Bastian Jäckl[1], Vojtěch Kloda[2], Daniel A. Keim[1], Jakub Lokoč[2]
1. Data Analysis and Visualization Group, University of Konstanz, Germany
2. SIRET Research Group, Charles University, Czechia

October 01-03, 2025, Reykjavík, Iceland