

Universiteit Leiden

Master Computer Science

Where Do Video LVLMs Look? Analyzing Perturbation-Based Attribution for Video LVLMs

Name:	Rutger Berger
Student ID:	2498278
Date:	01/07/2026
Specialisation:	Artificial Intelligence
1st supervisor:	Hazel Doughty
2nd supervisor:	Daan Pelt

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CA Leiden
The Netherlands

Abstract

While recent large vision-language model architectures are increasingly advanced in video understanding, their inner mechanics remain largely a black box. Explainable AI methods attempt to make decisions of such models interpretable and understandable to human end-users. In this study, we address an important gap in XAI literature to show how generative LVLMs come to their conclusions when answering questions about videos. We propose a novel optimization-based perturbation method that extends current frameworks to video LVLMs. By clustering input over spatiotemporal tubelets and including an evolutionary optimizer with a hierarchical strategy, our method is able to provide localized heatmaps for videos of moderate complexity. We find however, that the perturbation objective is fundamentally ill-posed for this setting as it allows for multiple near-optimal heatmaps that do not necessarily align in identified regions. Our findings suggest the need for methods beyond standard perturbation concepts to faithfully explain LVLM outputs for videos. Code is publicly available at https://github.com/rutgerberger/UncoveringVideoVLMs.

Contents

1 Introduction **3**

2 Related Work **5**
- 2.1 Large Vision Language Models . . . 5
- 2.2 Heatmap Visualization Methods . . . 6
- 2.3 Applicability of Post-Hoc XAI . . . 8

3 Method **8**
- 3.1 Problem Setup . . . 9
- 3.2 Mask Formulation . . . 10
- 3.3 Algorithm . . . 11
  - 3.3.1 Search Space . . . 11
  - 3.3.2 Evolution Strategies . . . 12

4 Experiments **15**
- 4.1 Data . . . 15
- 4.2 Experimental Setup . . . 16
- 4.3 Evaluation . . . 17

5 Results **18**
- 5.1 Hyperparameter and Ablation Study . . . 18
- 5.2 Solution Consistency . . . 22
- 5.3 Video Complexities . . . 24
- 5.4 Comparison . . . 26

6 Discussion **28**
- 6.1 Limitations . . . 29
- 6.2 Future Directions . . . 30

7 Conclusion **30**

8 Acknowledgements **31**

A Theoretical Analysis **36**

B Complete Experimental Setup **38**

C Full Hyperparameter Study **39**

D Additional Qualitative Results **40**

1 Introduction

In recent years, large vision language models (LVLMs) have emerged as state-of-the-art architectures for artificial visual understanding. Temporal modeling over long-term dependencies of high-dimensional video input has been a longstanding gap in literature which recent architectures successfully bridge [5, 35, 48], being able to answer questions about videos that take up to multiple tens of minutes [6]. Despite their remarkable success, the black-box nature of LVLMs poses significant challenges in critical domains, such as medical imaging, robotics or autonomous driving [4, 14]. For such applications, it is considered desirable that domain experts can verify outputs to confirm that they are grounded in relevant evidence.

Explainable AI (XAI) methods attempt to mitigate the black-box nature of such machine learning models. Generally, these methods should increase trust and understanding for end users who may have limited technical background, by explaining how models reach their final decisions. Notably, there are ongoing issues in the field of XAI. For instance, the distinction between an explanation's faithfulness and its plausibility remains debated [2, 20]. Because an explanation must simplify the underlying model to be interpretable, it inherently functions as an approximation of the model's internal causal rules. Methods may generate heatmaps that are visually appealing to humans but fail to reflect the actual model's decision process. More problematically, different XAI methods often lead to different explanations for the same decision [26]. The mathematical objectives and assumptions underlying these methods therefore require careful evaluation, especially when scaling from images towards video data.

Perturbation methods form an important subdomain of XAI [8, 24]. These methods attempt to quantify the importance of input features by analysing the effect of feature perturbations on the output score. Early works like RISE [37] or SHAP [34] approximate importance by sampling random perturbations and modeling the perturbation-output relationship. These approaches often give imprecise results and are computationally heavy. Optimization-based perturbation methods instead optimize a mask so that the model is no longer able to make a prediction, i.e., they optimize a mask to minimize the model's certainty, often constrained by a regularizer to enforce compactness [13, 39]. Recently, iGOS++ [25] proposes a novel optimization objective that pushes for a perturbation that correctly removes evidence when hidden (deletion), but also restores confidence when shown in isolation (insertion), achieving state-of-the-art explainability metrics across a range of image benchmarks. However, an important gap remains as there exists, to the best of our knowledge, no foundational method for perturbation-based XAI to deal with video LVLMs.

Moving from images to videos makes optimization-based perturbation substantially more difficult. Not only does the search space grow, the fact that evidence is distributed across time creates complex temporal dependencies. A region that appears unimportant in one frame may be essential when viewed in the context with surrounding frames. Similarly, masked evidence can sometimes easily be reconstructed by temporal redundancies and masking essential features may have limited effect. This causes flat optimization landscapes: obscuring an object in frame t will make the model use frame $t \pm 1$ and the output score will not change. As a result, finding the most salient perturbations requires

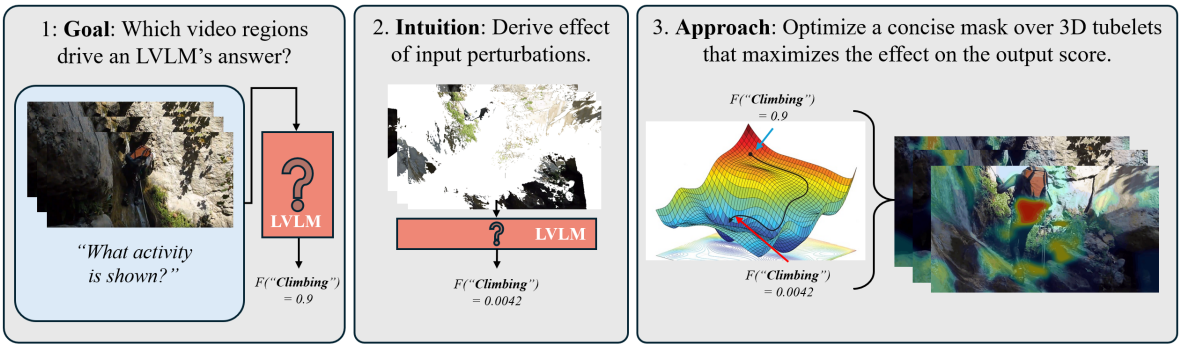


Figure 1: An intuitive overview of our approach. We perturb the input video by blending the video with a baseline, which can be a constant or blurred version of the video, ruled by the values of the mask. The goal is to find a concise mask that maximally affects the model's certainty, which is in practice done by optimizing the mask over a set of spatiotemporal tubelets. Final saliency is derived according to the found mask values, which in turn correspond to the (in)visibility of each tubelet.

an optimization over a high-dimensional and context-dependent landscape. Standard image-based perturbation methods fail as they optimize spatial masks independently per frame on a pixel-basis. They suffer from the curse of dimensionality and do not account for recurrent spatiotemporal concepts in videos [30, 54].

In this work, we propose a novel method to capture spatiotemporal evidence and optimize masks in a way that is aligned with video reasoning. Our goal is to create faithful heatmaps over input videos to reveal what features mostly contribute to LVLMs when answering questions about videos. We adjust the optimization landscape by clustering input over *tubelets* that align with input features over time. Clustering features into tubelets reduces the dimensionality of the search space and accounts for redundant information across frames, as optimization can then capture coherent concepts throughout the video. Since autoregressive output may depend on the visual input to varying degrees, we adopt the keyword selection framework of Xing et al. [56] to identify output tokens grounded in the visual input as optimization target. Subsequently, to deal with high-dimensional and non-convex optimization landscapes for video perturbations, we construct the mask using Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [16], a stochastic and derivative-free optimization method for non-convex continuous optimization problems. We maintain the perturbation objective from previous work [25, 39] and optimize masks for both deletion and insertion. Figure 1 shows an intuitive overview of our approach.

Our results show that, across datasets, our generated heatmaps are qualitatively commonly more localized compared to previous methods, such as iGOS++ [25], when applied to videos. However, across all ablations over our methodological advancements, none of the proposed improvements reached on average convergence towards an optimal solution for the perturbation objective. Our theoretical and experimental analyses suggest that the objective that is standard for perturbation-based methods is ill-posed, not only when applied to video LVLMs. A central issue is that the objective allows for multiple near-optimal solutions that do not necessarily share identified regions. This

questions the existence of a single, unified explanation. Moreover, assuming an ideal optimizer could locate the absolute mathematical minimum, this optimum does not necessarily align with true semantic relevance, as the objective does not differentiate between model confusion due to out-of-distribution (OOD) artifacts and the removal of actual visual evidence during optimization. Similarly, the non-convex landscape hinders finding these theoretical optima, as caused by, among others, asymmetry between deletion and insertion objectives and the high non-linearity of the LVLM. Concretely, our results show that perturbation-based derived heatmaps will at most show an approximation of visual evidence for videos.

In this study, we research how we can extend optimization-based perturbation methods from image towards video LVLMs. To this end, we investigate the following main research questions: **(1)** How do different design choices, such as clustering strategies and regularization, affect the convergence to the perturbation objective? **(2)** Does the optimization landscape allow for consistent and unique explanations across independent runs? **(3)** How does our method scale over data of increased complexity compared to existing baselines?

To summarize, our main contributions are:

- **A Novel Video Attribution Framework.** We introduce a hierarchical optimization-based perturbation method for video LVLMs. By combining CMA-ES with 3D spatiotemporal tubelets, our approach addresses the non-convex search space that complicates gradient-based optimization in videos.
- **Evaluation of Temporal Reasoning.** We categorize video question-answering tasks using the CLEVRER dataset and isolate different aspects of temporal reasoning. We show that question types vary in how evidence is distributed, which makes it difficult for the joint perturbation objective to recover a complete and faithful mask consistently.
- **Analysis of the Perturbation Objective.** We show that the standard joint objective is ill-posed for video input, as different runs converge to distinct near-optimal masks that trade off coverage of evidence, conciseness, and perturbation terms differently.

The remainder of this thesis is structured as follows. Section 2 details related work to our method and Section 3 outlines our method itself. Section 4 highlights the experimental setup, after which the results are shown and analyzed in Section 5. We end with a discussion and conclusion in Section 6 and 7, respectively.

2 Related Work

2.1 Large Vision Language Models

Essentially, large vision-language models (LVLMs) connect visual encoders (e.g., a Vision Transformer (ViT) [10]) with large language models (LLMs) and reach state-of-the-art multi-modal understanding [6, 35, 48]. In a typical architecture, visual inputs are split

up in squared patches which are then encoded into token representations and aligned with the LLM embedding space through a modality connector, such as cross-attention (Flamingo [3]) or learned projection layers (BLIP-2 [28], LLaVA [32]). Extending these models to video introduces additional challenges due to the high-dimensional spatiotemporal input, which are commonly addressed through frame sampling, token compression, or temporal aggregation strategies (e.g., VideoMAE [11, 50], Video-LLaVA [31]). In this study, we adopt two LVLM architectures with distinct designs: Video-LLaVA [31] and Qwen2.5-VL [5]. Both follow the current standard architecture of a vision encoder, a cross-modal projector, and an LLM, but differ in spatiotemporal modeling. In short, Video-LLaVA encodes frames separately and relies on the LLM to model temporal relations, while Qwen2.5-VL uses windowed self-attention and grouped token compression to more efficiently to model spatiotemporal structure more efficiently.

2.2 Heatmap Visualization Methods

Two prominent lines of research to obtain heatmaps that reflect model's evidence are *gradient-* and *perturbation-based* methods. We consider most relevant works to our method in each line here.

Gradient-Based Approaches Gradient-based methods estimate input influence by backpropagating the function over the input with respect to the output [9, 43, 45, 46, 58, 59]. While popular, there is no actual mathematical proof for the correctness of such approaches [53]. A core limitation of early methods is the saturation of gradients: if a model is entirely certain of its prediction, the local gradient becomes zero even if the feature is essential. *Integrated Gradients (IG)* [7, 47] solves this through calculating attribution by integrating local gradients along a linear path from an obscured baseline (e.g., a completely black image) to the original input. This ensures essential features are captured before the model saturates. However, this mathematical formulation comes with a strict constraint: it relies on extracting reliable and informative gradients across a continuous path, which is in practice approximated by the Riemann sum. Approximating reliable gradients for videos becomes unstable, due to recurrent spatiotemporal concepts which increase contextual dependencies of gradients and a significantly increased dimensionality. We address this issue by moving to a derivative-free perturbation strategy.

Perturbation-Based Attribution Perturbation methods measure how input modifications affect prediction certainty [34, 37, 40]. Optimization-based variants reframe this as finding the minimal mask that destroys the model's predictive confidence, originally proposed by *Meaningful Perturbations* [12, 13]. In this line, iGOS [39] combines mask optimization with Integrated Gradients [47] and estimate mask importance by integrating gradients along a path from an obscured baseline $\tilde{I}$ to the masked image $\Phi(I, \tilde{I}, M)$, as given by $\nabla_I^{IG} f(M) = \frac{1}{S} \sum_{s=1}^{S} \frac{\partial f\left(\Phi\left(I, \tilde{I}, \frac{s}{S} M\right)\right)}{\partial M}$. Here Φ perturbs the baseline input I with the obscured baseline $\tilde{I}$ by element-wise multiplication, guided by mask M. Gradients are derived of the mask values with respect to the model's output score,

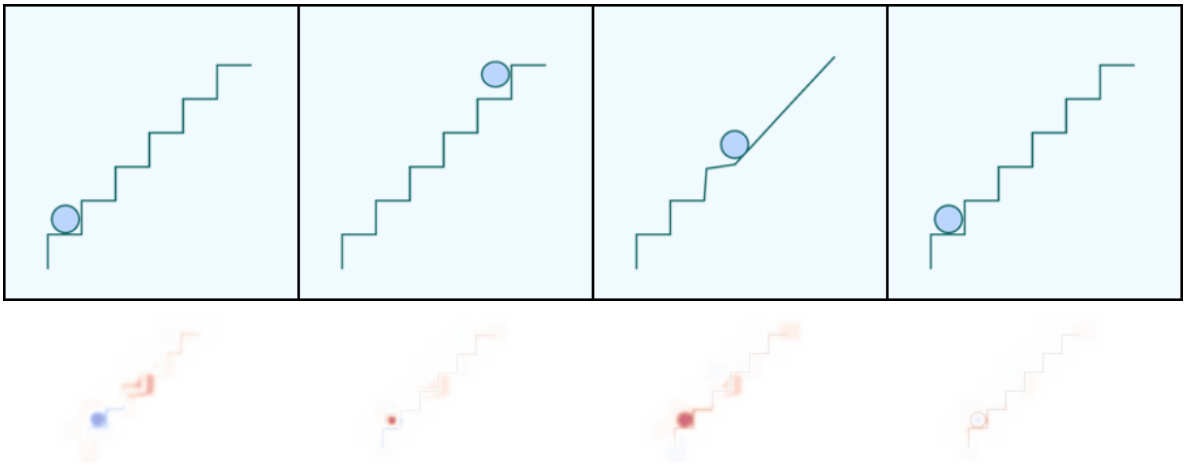


Figure 2: Contradictory gradients observed along the Integrated Gradients (IG) path when applying the method of Xing et al. [56]. Top: The original input video, with the question 'What color is the central object in this video?' (Answer: 'Blue'). Bottom: Gradients computed for the first frame at successive interpolation steps along the IG path, transitioning from a fully obscured baseline (leftmost) to the unmodified video (rightmost). Note that the gradients at the salient central object are initially negative but invert to positive values by the end of the path, highlighting the instability of gradient-based attribution in this context

pushing the mask to minimize the model's certainty. Subsequently, iGOS++ [25] found introducing some minor perturbations alone can easily hurt the confidence without aligning to all evidence in the input. Therefore, apart from optimizing a mask M_x for *deletion*, iGOS++ optimizes a separate mask M_y simultaneously for *insertion* to restore confidence by only inserting the masked regions. The Hadamard product of both ($M = M_x \odot M_y$) is used as final mask. Recently, Xing et al. [56] extend the iGOS framework to LVLMs by incorporating a non-convex search method and introducing a keyword filtering strategy to adapt the masking objective to free-form generative text.

While successful for images, including IG breaks down for video. The addition of the temporal dimension complicates the search space and we find that gradients can become contradictory along the path required by IG, as shown in Figure 2. The used Riemann sum becomes in this case insufficient to approximate the entire continuous path and gradients may sum each other out. In this work, we include the insertion and deletion optimization objectives to construct a mask but discard the IG framework by optimizing masks using Evolutionary Strategies (ES), an optimizer that is suited for the high dimensional non-convex video landscape.

Video Attribution Methods Video attribution is harder than image attribution because explanations must account for both spatial and temporal dependencies. Direct adaptations of image methods often produce noisy or unstable maps [30, 52, 54]. SWAG-V [17, 18] collects gradients in a single backward pass and overcomes uninformative gradients by averaging them across 3D supervoxels. STAA [54] uses attention weights to assign relevance to regions, while Li et al. [30] extend image-based perturbation to the

video domain by optimizing a 3D mask over the entire video on pixel-level, regularized by a 3D-convolutional loss function to promote spatial and temporal smoothness. Uchiyama et al. [51, 52] adopt an object-tracking framework to propagate spatial masks along motion trajectories to maintain temporal consistency. Similarly, STCE [21] groups pixels into cohesive spatiotemporal tubelets and apply perturbations to isolate the impact of distinct moving concepts on the final classification. While successful, each of these methods are explicitly designed for either 3D-CNNs or ViTs discriminative models.

When scaling from discriminative methods toward LVLMs, issues arise in the sense that the attribution target in an LVLM is ambiguous. Specifically, discriminative methods optimize for a static function $f(x)$, whereas LVLMs model an autoregressive joint probability $\prod_{t=1}^{K} p(y_t \mid y_{<t}, x)$ and linguistic priors complicate this target. Moreover, discretising visual features into text tokens and their alignment via cross-modal projection layers significantly complicate the input-output relationship.

Our method addresses this gap between discriminative video XAI and generative LVLMs. To deal with autoregressive models for video processing, we unify the keyword selection strategy of Xing et al. [56] to select faithful targets based in visual evidence with the intuition of prior work to cluster input in tubelets across time and space. Subsequently, we adjust to the increased non-linearity as caused by the LVLM by optimizing using CMA-ES [16], an adaptation to ES designed for increasingly non-linear search spaces.

2.3 Applicability of Post-Hoc XAI

Explainable AI methods are useful, but they come with known limitations [22, 42, 55]. A central issue is the faithfulness trade-off: an explanation that fully reproduces the model would no longer be simpler than the model itself [2]. This limitation is especially relevant for post-hoc methods, which must simplify the underlying decision process. Moreover, in many works explanations are posed as method to improve trust ([37, 40]), but they can also mislead users when they appear plausible without being faithful [38]. In a recent review by Rong et al. [41] it has been found that only half of the human studies into the effect of XAI on human trust finds positive outcomes [22]. Similarly, it is a common misperception that we need to fully understand a system in order to trust it [2] (think about the human brain). We therefore explicitly treat produced heatmaps as approximations of model behavior and do not present them as method to foster user trust. To mitigate the risk of generating explanations that seem visually plausible but are unfaithful to the model, we evaluate with complementary quantitative metrics that present the reader a complete image of the faithfulness of our heatmaps. Moreover, we analyse the perturbation objective both theoretically and experimentally to investigate whether it allows for valid explanations.

3 Method

Our goal is to find the most influential regions of input videos for an LVLM's output. In practice, this translates to finding the input features that confuse the model when hidden

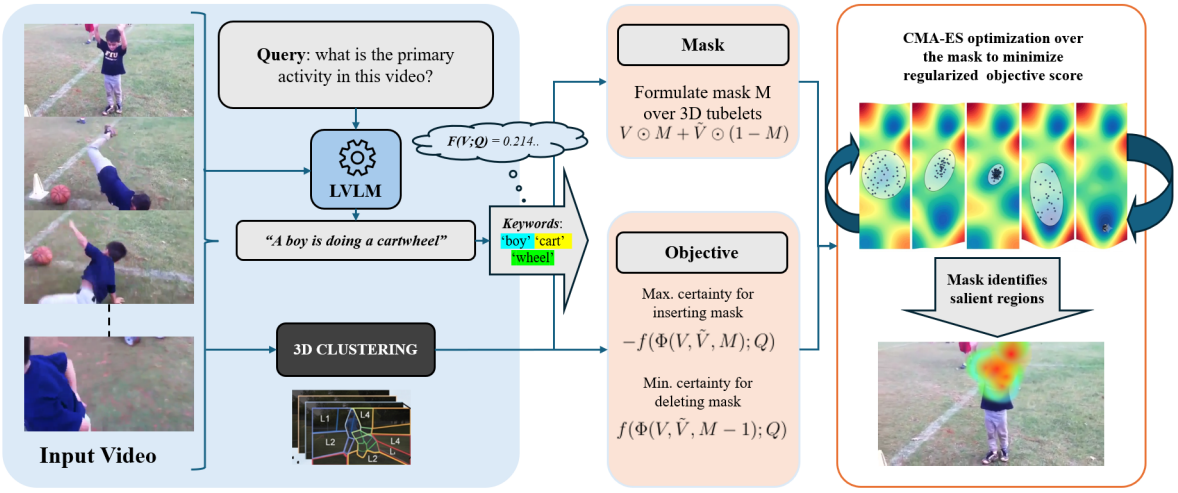


Figure 3: A schematic overview of our method. The mask M perturbs the given input video V with a baseline video $\tilde{V}$ by Φ (Eq. 1, shown in the upper box 'Mask'). The key intuition behind our method is that we optimize this mask M so that it covers all regions that constitute to the model's certainty, while remaining concise and connected. Creating 3D tubelets groups the search space into coherent spatiotemporal concepts. CMA-ES employs an evolutionary approach which adaptively transforms the search distribution of candidates so it overcomes local minima, often given by highly local but insufficient evidence, or noisy artifacts.

and reconstruct evidence when shown. We formulate this into a mask optimization problem, in which we optimize a mask M for both *deletion* and *insertion*. We construct the mask over spatiotemporal clusters (*tubelets*) and optimize using Evolutionary Strategies to maximally affect the model's decisions. A summary of our approach can be found in Figure 3.

3.1 Problem Setup

We follow perturbation-based methods by optimizing a continuous mask $M \in [0, 1]$ that modifies the original video V through element-wise multiplication [13, 39]. The masking operation Φ perturbs the original video V by interpolating it with an uninformative baseline $\tilde{V}$, given by:

$$\Phi(V, \tilde{V}, M) = V \odot M + \tilde{V} \odot (1 - M) \tag{1}$$

where $\odot$ denotes the Hadamard product. Under this formulation, $M = \mathbf{1}$ represents the original video V and $M = \mathbf{0}$ completely destroys all evidence with $\tilde{V}$. In practice $\tilde{V}$ can be a blurred version of the video, noise or a constant. To capture the model's certainty, we let $f(V; Q)$ be a short-hand notation for the output score of a target answer a given the video V and text query Q. For an autoregressive model, this score is the log-likelihood of the answer sequence of T tokens, $a = (a_1, a_2, \ldots, a_T)$, obtained via:

$$f(V; Q) = \log P(a \mid V, Q; \theta) = \sum_{t=1}^{T} \log P(a_t \mid a_{<t}, V, Q; \theta) \tag{2}$$

Here, P represents a conditional probability distribution parameterized by the VLM's weights θ. It defines a categorical distribution over the vocabulary $\mathcal{V}$, mapping the given video, query, and preceding tokens to a probability:

$$\sum_{w \in \mathcal{V}} P(a_t = w \mid a_{<t}, V, Q; \theta) = 1 \tag{3}$$

In this study, we adopt two LVLM architectures with distinct designs: Video-LLaVA [31] and Qwen2.5-VL [5]. Both follow the current standard architecture of a vision encoder, a cross-modal projector, and an LLM, but differ in spatiotemporal modeling. Video-LLaVA represents a video as a sequence of frames $\{I_t\}_{t=1}^{T}$, each encoded into patch tokens $Z_t \in \mathbb{R}^{N \times d}$, projected into the LLM space, and concatenated as $Z = [Z_1, \ldots, Z_T]$. The resulting sequence of length $T \cdot N$ is processed autoregressively, with temporal dependencies implicitly modeled by the LLM. In contrast, Qwen2.5-VL employs a modified ViT with windowed self-attention, reducing complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, and an MLP-based projector which compresses *grouped* patch tokens. It also introduces Multimodal Rotary Position Embeddings with $p = (t, h, w)$, where the temporal index is aligned to absolute time.

3.2 Mask Formulation

Recall that we optimize the mask M for two goals: *deletion* (necessity) and *insertion* (sufficiency). For deletion, we want to minimize the output score f by removing evidence. This is captured by minimizing $f(\Phi(V, \tilde{V}, 1 - M); Q)$. By using $1 - M$, we invert the relationship from Equation 1, and as $M \to \mathbf{1}$, we destroy all evidence by replacing the video with $\tilde{V}$. For insertion, we want to maximize f by providing sufficient evidence to restore the model's confidence when only the masked regions are shown. To capture this, we maximize $f(\Phi(V, \tilde{V}, M); Q)$, where inserting V for $\tilde{V}$ as $M \to \mathbf{1}$ introduces visual evidence and improves the model's confidence.

Naively combining both objectives induces no constraints on the size of the mask. Moreover, it promotes sensitivity to noisy and flickering masks which artificially reduce confidence but have no causal relationship to real-world evidence [21, 39]. For this reason, we employ L1 regularization on the mask size ($\|M\|_1$) and total variation regularization to promote spatial and temporal smoothness. We trade-off the strength of this regularization by a scalar λ_{reg} and denote the regularization by $g(M) = \|M\|_1 + \mathrm{TV}(M)$. Combining insertion, deletion and regularization terms, we formalize our *joint* objective as

$$\min_{M} \left[-f(\Phi(V, \tilde{V}, M); Q) + f(\Phi(V, \tilde{V}, 1 - M); Q) + \lambda_{reg} \cdot g(M) \right], \tag{4}$$

where we optimize a mask to obscure evidence when applied and introduce evidence when shown.

Separate Objective As Equation 4 defines a *joint* objective for insertion and deletion, either insertion or deletion may easily dominate throughout stages, which can hinder convergence and lead to local optima.

An underlying issue is that the joint objective does not distinguish between necessity and sufficiency, so optimization may converge on a partial mask that explains the prediction only incompletely. Therefore, we also propose a *separate* optimization strategy, optimizing two masks over two independent runs. Intuitively, the separate optimization strategy overcomes stagnation of one objective while the other continues to converge by independently optimizing an insertion mask M_{ins} to maximize sufficiency, and a deletion mask M_{del} to maximize necessity:

$$M^*_{ins} = \arg\min_M \Big[-f(\Phi(V, \tilde{V}, M); Q) + \lambda_{reg} \cdot g(M)\Big] \tag{5}$$

$$M^*_{del} = \arg\min_M \Big[f(\Phi(V, \tilde{V}, 1 - M); Q) + \lambda_{reg} \cdot g(M)\Big] \tag{6}$$

The final saliency mask is then obtained via the Hadamard product, $M^* = M^*_{ins} \odot M^*_{del}$.

3.3 Algorithm

Having defined the optimization objective (Eq. 4 or 5 & 6), we want to find an optimal mask M^* that satisfies this minimization problem.

3.3.1 Search Space

Spatiotemporal clustering We cluster the input space into a set of spatiotemporal *tubelets* $\mathcal{T}$ using 3D Simple Linear Iterative Clustering (SLIC) [1], as supported by recent works in video XAI [17, 43]. While intuitively, forming spatiotemporal (3D) clusters is beneficial, it is also possible to apply SLIC on a per-frame basis, creating separate spatial and unconnected (2D) clusters for each distinct frame. We ablate this setting within our experiments to reveal the trade-off on performance.

Subsequently, we define a continuous weight vector $x \in \mathbb{R}^{|\mathcal{T}|}$ over the set of tubelets, and apply a sigmoid activation to restrict weights to $\sigma(x) \in [0, 1]$. The final video mask $M \in [0, 1]^{T \times H \times W}$ is constructed by mapping each tubelet's weight to its spatial-temporal coordinates:

$$M_{(t,h,w)} = \sum_{\tau \in \mathcal{T}} \sigma(x_\tau) \cdot \mathbb{I}_{(t,h,w) \in \tau} \tag{7}$$

Where $\mathbb{I}_{(t,h,w)\in\tau}$ is an indicator function that equals 1 if the spatial-temporal coordinate (t, h, w) belongs to tubelet τ, and 0 otherwise. Consequently, we directly optimize M via weights $\mathbf{w}$. By applying a sigmoid function we map the weights to the $[0, 1]$ interval without creating flat plateaus so the algorithm retains a direction to follow at boundaries.

Keyword Selection In autoregressive generation, not every token is dependent on the same amount of visual information. To make sure the mask is being driven by actual visual evidence, we follow the work of Xing et al. [56] to identify tokens that depend on the input video and not on the preceding language. We do this by comparing the model's confidence under the original video V with its confidence under the uninformative

baseline $\tilde{V}$. For each token a_t, we measure this difference with the Log-Likelihood Ratio (LLR):

$$LLR_t = \log P(a_t \mid a_{<t}, V, Q; \theta) - \log P(a_t \mid a_{<t}, \tilde{V}, Q; \theta) \tag{8}$$

A large LLR_t indicates that a token is visually grounded, since its probability decreases strongly when the video is replaced by the baseline. Tokens with little visual dependence instead will show only minor change. We therefore define the set of visually relevant tokens $\mathcal{K}$ over the set of tokens which LLR_t exceeds a threshold α, while excluding the first generated token ($t = 1$), as this often determines sentence structure instead of fine-grained visual content [56]:

$$\mathcal{K} = \{a_t \mid LLR_t > \alpha, t \neq 1\} \tag{9}$$

Note that we now redefine the score $f(V; Q)$ so that it only aggregates over the visually relevant tokens:

$$f(V; Q) = \sum_{a_t \in \mathcal{K}} \log P(a_t \mid a_{<t}, V, Q) \tag{10}$$

While standard Log-Likelihood Ratios are computed post-softmax (as a probability distribution is assumed in Eq. 8), we find the softmax squashes probabilities. The relative probabilities may remain stable even as the model's raw internal confidence decreases. To address this, we compute $f(V; Q)$ and LLR_t using the *pre-softmax logits*, which correspond to the raw network output of the LVLM. By optimizing directly on the raw logits, the objective landscape becomes more responsive to perturbations that affect the model's true internal evidence.

3.3.2 Evolution Strategies

Having the search space defined, we want to optimize the continuous tubelet weights x over the non-convex landscape of the selected keywords of the LVLM's raw generative output tokens. To this end, we adopt Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [16]. Our implementation bases itself on that of Hansen et al. [15]. Essentially, CMA-ES iteratively samples a population of candidates from a multivariate normal distribution and adaptively updates this distribution based on the scores over this generation. Commonly, this algorithm runs for a fixed number of generations G. We now provide a concise outline of each step and parameter for CMA-ES.

The algorithm samples a population of λ candidate solutions from a multivariate normal distribution at each generation g:

$$x_k^{(g+1)} \sim m^{(g)} + \sigma^{(g)} \mathcal{N}(0, C^{(g)}) \quad \text{for } k = 1, \ldots, \lambda \tag{11}$$

where $x_k^{(g+1)}$ represents a candidate tubelet weight vector, $m^{(g)} \in \mathbb{R}^{|\mathcal{T}|}$ is the current mean vector of the search distribution, $\sigma^{(g)} \in \mathbb{R}_{>0}$ is the global step-size, and $C^{(g)} \in \mathbb{R}^{|\mathcal{T}| \times |\mathcal{T}|}$ is a symmetric and positive definite covariance matrix. After evaluating the candidate solutions and sorting them by their objective function values, the algorithm updates the mean vector of the search distribution $m^{(g+1)}$ by a weighted sum of the μ best-performing candidate vectors $x_{i:\lambda}^{(g)}$, using recombination weights ω_i (static values with $\sum \omega_i = 1$):

$$m^{(g+1)} = \sum_{i=1}^{\mu} \omega_i x_{i:\lambda}^{(g+1)}$$

The key intuition behind this algorithm is to keep track of an *evolution path*. If subsequent generations are consistently moving towards the same direction, we want to take larger steps towards that direction. Similarly, if the path is unaligned, we decrease the step-size.

First, p_σ adapts the global step size σ and intuitively captures the vector sum of the preceding evolution steps taken. If the recent mean vectors align, p_σ is a relatively larger vector and we want to increase the step-size. Instead of using a rolling window where all previous generations are equally important, we update p_σ by using exponential smoothing, using learning rate $c_\sigma \in [0, 1]$, the inverse square root of the covariance matrix $C^{(g)}$ and the previous and current mean vectors m:

$$p_\sigma^{(g+1)} = (1 - c_\sigma)p_\sigma^{(g)} + \frac{\sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}}}{\sigma^{(g)}}(C^{(g)})^{-1/2}(m^{(g+1)} - m^{(g)})$$

Here, $\mu_{\text{eff}} = \frac{1}{\sum_{i=1}^{\mu} \omega_i^2}$ quantifies the effective number of parents contributing to the new search distribution. If μ_{eff} is large, the new search direction $m^{(g+1)}$ was formed by a broad consensus of many high-performing candidates (low variance, high reliability). The term $\sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}}$ cancels out variance accumulated across the samples. Multiplying by $(C^{(g)})^{-1/2}$ normalizes the vector space back into a perfect sphere, as the covariance matrix C might be stretched into a long ellipsoid shape. Taking a big step along the length of the ellipsoid may be normal for the stretched landscape. If $\|p_\sigma\|$ is larger than a random walk expectation, global step size σ goes up as the updates are consistently pushing into a unified direction:

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p_\sigma^{(g+1)}\|}{\mathbb{E}\|\mathcal{N}(0, \mathbf{I})\|} - 1\right)\right) \tag{12}$$

Here, d_σ is the 'damping parameter' to scale the updates accordingly.

To adaptively update the covariance matrix C, the algorithm essentially takes the weighted average of three terms. We maintain a separate evolution path p_c which is roughly similar to p_σ but is unnormalized, as we want to include direction. Boolean variable $h_\sigma \in \{0, 1\}$ analyses the squared norm of p_σ and prevents the p_c from updating when the magnitude of p_σ is too large. Path p_c incorporates this safeguard alongside its own learning rate c_c:

$$p_c^{(g+1)} = (1 - c_c)p_c^{(g)} + h_\sigma \frac{\sqrt{c_c(2 - c_c)\mu_{\text{eff}}}}{\sigma^{(g)}}(m^{(g+1)} - m^{(g)})$$

The covariance matrix $C^{(g+1)}$ is then constructed by combining the decayed previous matrix, a rank-one update from p_c, and a rank-μ update derived from the outer products of the current generation's search steps. Using c_1 and c_μ as learning rates, and a modified rank-one decay term $c_{1a} = c_1(1 - (1 - h_\sigma)c_c(2 - c_c))$ to adjust for variance loss when h_σ

Algorithm 1 Overview of Steps using CMA-ES Optimization.

Require: Video V, Question Q, Model $\mathcal{M}$, Baseline $\tilde{V}$, Generations G
Ensure: Saliency Mask M^*

$A \leftarrow \mathcal{M}(V, Q)$ {Generate initial response}
$\mathcal{K} \leftarrow \{a_t \in A \setminus \{a_1\} \mid LLR_t > \alpha\}$ {Visually grounded tokens via Eq. 8 & 9}
$\mathcal{T} \leftarrow \text{SLIC}(V)$ {Form tubelets / clusters}
$\mathcal{S} \leftarrow$ Set of super clusters over $\mathcal{T}$
Initialize CMA-ES parameters: $m^{(0)} \in \mathbb{R}^{|\mathcal{S}|}$, $\sigma^{(0)}$, $C^{(0)} \leftarrow \mathbf{I}$, population size λ
for generation $g = 0$ **to** $G - 1$ **do**
  if $g = G/2$ **then**
    $m^{(g)} \leftarrow$ Unpack $m^{(g)}$ over $\mathcal{T}$ {Transition to fine-grained tubelets}
    Reinitialize parameters: $\sigma^{(g)} \leftarrow \sigma^{(0)}$, $C^{(g)} \leftarrow \mathbf{I}$
  end if
  Sample solution vectors $x_k \sim m^{(g)} + \sigma^{(g)}\mathcal{N}(0, C^{(g)})$ for $k = 1, \ldots, \lambda$
  for each candidate solution x_k **do**
    Generate mask M_k from x_k over $\mathcal{T}$ via Eq. 7
    Evaluate fitness $\mathcal{F}_k$ over $\mathcal{K}$ using Eq. 4
  end for
  Update parameters $m^{(g+1)}, \sigma^{(g+1)}, C^{(g+1)} \leftarrow \text{CMA-ES}(x_{1:\lambda}, \mathcal{F}_{1:\lambda})$
end for
return M^* {Return mask of best individual x_k via Eq. 7}

is triggered, the complete covariance update is formulated as:

$$C^{(g+1)} = (1 - c_{1a} - c_\mu \sum \omega_k) C^{(g)} + c_1 p_c^{(g+1)} (p_c^{(g+1)})^T + c_\mu \sum_{k=1}^{\lambda} \omega_k \left(\frac{x_{k:\lambda}^{(g+1)} - m^{(g)}}{\sigma^{(g)}} \right) \left(\frac{x_{k:\lambda}^{(g+1)} - m^{(g)}}{\sigma^{(g)}} \right)^T \tag{13}$$

Note that we use $\sigma^{(g)}$ and $m^{(g)}$ and not these of the new generation.

Hierarchical Optimization. We introduce a hierarchical optimization strategy to overcome converging to local minima early on. Instead of optimizing the mask weights x over all tubelets from the beginning, we first group them into a smaller set of super-clusters using K-Means clustering based on spatial distance. The CMA-ES optimizer is initialized on this lower-dimensional space so we force the method to first locate the broader regions of interest. Algorithmically, the code executes two entirely separate CMA-ES loops. After an initial number of generations, we unpack the super-cluster weights and transfer the mean vector m to the fine-grained tubelets that belong to the associated super-clusters. However, the parameters of the learned distribution from the first stage are discarded (σ, C, evolution paths p_c, p_σ) as we operate on a different optimization landscape. Therefore, the hierarchical phase only locates the center of the search space to a promising coarse region, while the optimizer is forced relearn the local variance and directional relationships after switching to fine-grained optimization. This progressive approach prevents the optimizer from becoming stuck in local minima early in the search process. Full pseudocode is shown in Algorithm 1.

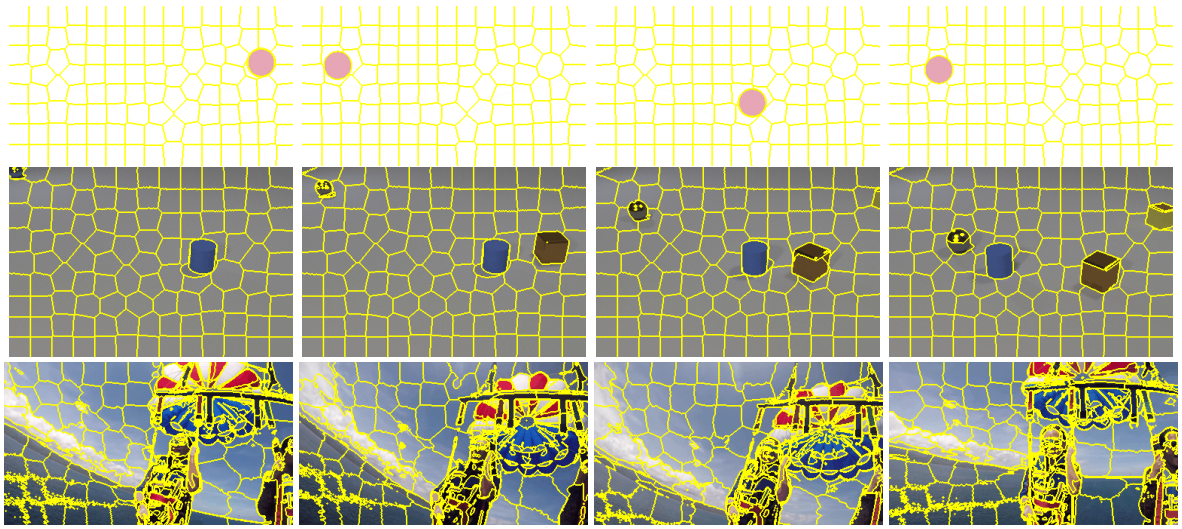


Figure 4: SLIC spatiotemporal clusters formed over example data from datasets Simple (above) vs. CLEVRER (middle) and K400 (below). Commonly, the number of clusters representing salient objects increases for CLEVRER and especially for K400. While the temporal aspect is hard to visualize, represented clusters are existent across frames. Observe that the clusters become noisy over the exact objects, failing to capture salient objects coherently.

In Appendix A, we provide a theoretical analysis on the optimization problem in Eq. 4. Concretely, moving to video perturbations for LVLM significantly increases optimization difficulty. More problematically, there is no theoretical guarantee that there is a unique solution and even if an optimum is found, it does not have to imply that this mask is aligning with actual semantic relevance.

4 Experiments

To evaluate our framework, we propose a series of experiments aimed at answering our three key research questions. Specifically, we test our algorithmic design choices (RQ1), evaluate the mathematical consistency of the optimization landscape (RQ2), and benchmark our method's scalability against existing baselines (RQ3). To answer these questions, we evaluate our method across four distinct datasets, using two models and three complementary evaluation metrics.

4.1 Data

An overview of included datasets can be found in Table 1. *Simple* is a custom synthetic dataset containing videos of only single moving objects. Questions explicitly do not contain any temporal reasoning and this serves as a baseline sanity check to verify whether our method is sensitive to evidence. The dataset is made publicly available

Table 1: Overview of included datasets. The limited number of included videos in experiments is due to high computational demands, as each individual experiment takes up to 2-3 hours.

Dataset	Purpose	Test Size
Simple	Qualitative sanity check.	40 videos (all)
CLEVRER [57]	Isolate aspects of question-answering for videos.	30 sampled videos (from 5,000)
Kinetics-400 [23]	Test method on real-world action data.	30 sampled videos (from 18,035)
CV-Bench [49]	Image baseline for additional experiments.	916 images

on HuggingFace[1]. *CLEVRER* [57] is a synthetic video dataset of increased complexity containing multi-object interactions and collisions. We include the validation split. Using provided program annotations, we categorize the data into six isolated aspects of temporal reasoning (Table 2) to observe what real-world video aspects complicate optimization complexity. *Kinetics-400* (K400) [23] is a common benchmark in video XAI [17, 54], representing real-world action video data each classified into its belonging action class. Our experiments run over the *validation* split which consists of 18,035 samples, on which we again take 30 samples due to high time complexities. With K400 we are able to compare the performance of our method to known records on real-world data. Finally, *CV-Bench* [49] is a real-world image dataset. Originally, Xing et al. [56] use this data to verify iGOS for LVLM on images. Therefore, we use this to expand our findings and verify optimization inconsistency for the perturbation objective. We evaluate on the subset identified by Xing et al.[2].

4.2 Experimental Setup

We use Video-LLaVA [31] (7B) on Simple and Qwen2.5-VL [5] (7B) on CLEVRER, K400 and CV-Bench. Example frames with generated clusters for each dataset can be found in Figure 4. Throughout experiments, we consistently compare using configuration †, which is identified by regularization of $\lambda_{reg} = 10$, CMA-ES iterations = 50, hierarchical pre-softmax optimization and 3D spatiotemporal clustering. All other parameters remain constant and can be found in Appendix B.

We compare against the method of Xing et al. [56], adapted to video by extending the bilateral TV-norm with temporal consistency and optimizing jointly across frames. We refer to this baseline as iGOS. Hyperparameters are adopted from the original work (lr=10, $L1$=1, $L2$=0.1, $L3$=10) and verified empiricall to align for given video data. As extra baseline, we compare against randomized heatmaps, where each tubelet gets assigned a uniformly sampled score between [0,1].

[1]https://huggingface.co/datasets/rutgerberger/VideoLVLMAnalysisEasy/
[2]https://huggingface.co/datasets/xiaoying0505/LVLM_Interpretation

Table 2: Categories constructed over CLEVRER [57] using program annotations. Each category isolates a distinct aspect of video question answering and forms a disjoint subset of questions.

Cat.	Q. Type	Size	Program Filter	Example Question	Test Objective
A	descriptive	7,386	`filter_stationary`	How many stationary cyan objects are there?	Object localization
B	descriptive	30,240	`filter_in`, `filter_out`, `filter_moving`	What material is the last object that enters the scene?	Localization of moving objects
C	descriptive	11,333	`filter_collision`, `get_col_partner`	Which object collided with the gray sphere?	Multi-object interactions
D	explanatory	8,488	`filter_ancestor`	Which of the following is responsible for the sphere's colliding with the cylinder?	Causal evidence localization
E	predictive	3,557	`unseen_events`	Will the red sphere hit the cylinder in the future?	Predictive visual reasoning
F	counterfactual	9,333	`get_counterfact`	What would happen if the blue cube were removed?	Counterfactual reasoning

4.3 Evaluation

Evaluating XAI remains challenging as there are no ground-truth explanations and there is no standardized framework [29, 60]. Common metrics can be misleading; for example, the *pointing game* [59] measures alignment with human annotations but not faithfulness to model reasoning, allowing even edge detectors to perform well [2, 25]. We therefore adopt three complementary evaluation metrics.

Sufficiency & Necessity We can assess how well the optimizer satisfies Equation 4 using sufficiency and necessity [54, 56]. First, output log-probabilities given by Equation 2 are normalized per video with local minimum $p_{baseline} = f(\tilde{V}; Q)$ and the theoretical LLR maximum of 0:

$$f_{norm}(p) = \frac{p - p_{baseline}}{0 - p_{baseline}}. \tag{14}$$

Sufficiency evaluates recovery of confidence and corresponds to the first part of our objective:

$$\text{Sufficiency} = 1 - \Big(f_{norm}(V;Q) - f_{norm}(\Phi(V,\tilde{V},M);Q)\Big). \tag{15}$$

A sufficiency near 1 indicates that, when inserted, the masked regions perfectly capture the evidence required to reconstruct confidence.

Necessity measures confidence drop when removing masked regions and corresponds to the second part of our objective:

$$\text{Necessity} = 1 - \Big(f_{norm}(\Phi(V,\tilde{V},1-M);Q) - f_{norm}(\tilde{V};Q)\Big) \tag{16}$$

Similarly, necessity scores near 1 indicate that, when removed, the detected salient regions perfectly destroy the required evidence.

Scores are clipped to [0,1], as normalized scores may become negative when the masked score is lower than for $\tilde{V}$. These metrics reflect mostly optimization success relative to the objective and not semantic interpretability. Note that for these metrics, masking everything ($M = \mathbf{1}$) would yield perfect scores.

Deletion and Insertion Subsequently, we evaluate whether high-saliency regions are truly influential using the deletion and insertion area under the curve (AUC) metrics [25, 44, 56]. Deletion progressively replaces salient pixels in V with $\tilde{V}$, based on the assigned scores in M; lower AUC indicates faster confidence drop. Insertion starts from $\tilde{V}$ and restores pixels and a higher AUC indicates most important evidence has gained highest saliency.

Jaccard's Similarity Finally, we measure solution consistency via intersection over union scores on the top-K% tubelets:

$$\text{IoU}(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{17}$$

For iGOS [56], we analyse this metric using the top-K% of pixels. An IoU score near 1.0 implies consistently similar identified salient regions, while when dropping to 0 the top regions are barely shared.

5 Results

5.1 Hyperparameter and Ablation Study

The first set of experiments analyses our design choices and shows whether our method is able to find masks that adhere to the minimization objective. We perform both an hyperparameter analysis and an ablation study to obtain a complete insight.

Regularization Terms. Table 3 confirms the effect of regularization, as decreasing regularization λ_{reg} increases sufficiency, while AUC metrics deteriorate. Mathematically, the goal of regularization terms is to enforce a smooth, concise and connected mask. With

Table 3: The effect of the regularization term λ_{reg} on CLEVRER. While a loose regularization promotes sufficiency, AUC metrics strongly get better as λ increases. Regularization forces the mask to assign highest saliency to the most important objects, at the cost of creating an increasingly incomplete mask. Within this setting, it is hard to provide masks that are both concise and complete, suggesting that these are competing objectives. With † we denote the default configuration shared across ablation groups.

Parameter	**Value**	**Sufficiency ↑**	**Necessity ↑**	**AUC Ins ↑**	**AUC Del ↓**
λ_{reg}	0	.729 ± .316	.432 ± .361	.585 ± .223	.636 ± .183
λ_{reg}	1	**.737** ± .304	.413 ± .347	.695 ± .198	.580 ± .225
λ_{reg}	5	.689 ± .292	.395 ± .354	.721 ± .200	.497 ± .230
λ_{reg}	$10^\dagger$	.695 ± .299	.430 ± .368	.721 ± .212	.494 ± .219
λ_{reg}	20	.642 ± .318	**.449** ± .351	**.733** ± .196	**.478** ± .246

a stronger regularization, the optimizer is forced to highlight only the most important regions. This leads to considerably better AUC metrics, as evidently important evidence will be inserted or deleted earlier on. However, this comes at a price of completeness of masks, as sufficiency scores decrease substantially. Interestingly, with little regularization necessity scores are still low and even with $\lambda_{reg} = 0$, the optimizer does not come near the approximate optimum given by $M = \mathbf{1}$, as both sufficiency and necessity do not approach 1. With no regularization, masking all evidence is arguably an easy identifiable global region of interest which the optimizer fails to find.

As masks fail to be both concise and complete, our results suggest these are inherently competing objectives. By allowing masks to capture distributed evidence, we trade-off for localized explanations. However, the incomplete scores for zero regularization indicate our optimization setup is also insufficient for finding optimal solutions, implying that finding distributed evidence requires navigating a difficult optimization landscape.

Table 4: Results for increasing the number of iterations for the evolutionary optimizer. The positive effect of having more iterations diminishes after roughly 50 iterations. Consequently, convergence towards optimal masks appears not to be limited by the given search budget. Again, † denotes the default configuration shared across ablation groups.

Parameter	**Value**	**Sufficiency ↑**	**Necessity ↑**	**AUC Ins ↑**	**AUC Del ↓**
Generations	20	**.700** ± .197	.431 ± .348	**.721** ± .172	.519 ± .197
Generations	$50^\dagger$	.695 ± .299	.430 ± .368	**.721** ± .212	**.494** ± .219
Generations	100	.589 ± .338	**.433** ± .361	.699 ± .189	.496 ± .242

Iterations. Table 4 shows that increasing the number of iterations does not improve performance and can even reduce sufficiency, from .700 at 20 iterations to .589 at 100 iterations. Optimization difficulty is therefore not mainly caused by insufficient search budget. Interestingly, AUC metrics remain roughly similar, hinting at an equally

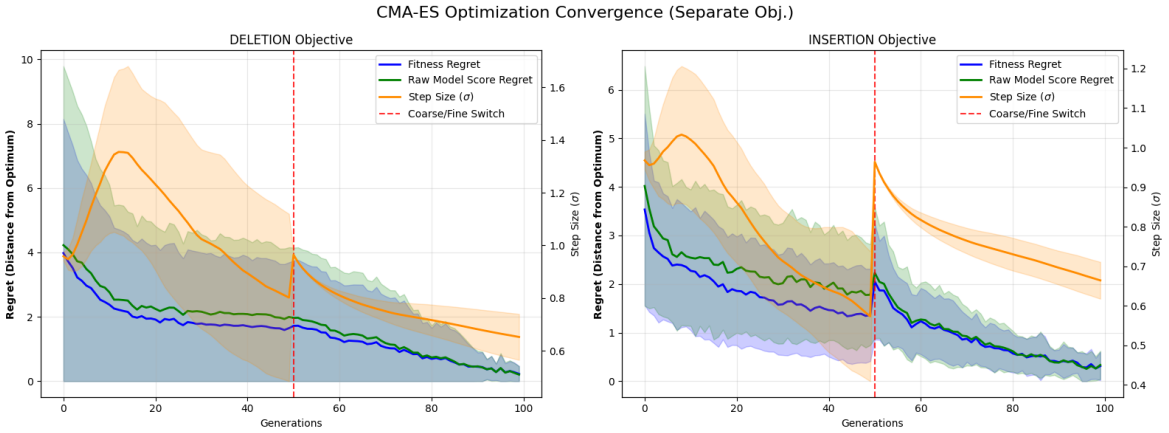


Figure 5: Averaged convergence plots for the separate optimization regime using 100 iterations (30 samples). The raw model score regret reflects the actual confidence of the model compared to the optimal confidence found during optimization (the minimum for deletion, maximum for insertion). Observe that during the first part of the optimization, the fitness is consistently lower compared to the raw model score. This difference is solely caused by the regularization term. Later on, this difference diminishes: the method bridges the regularization penalty.

localized set of evidence. Figure 5 provides the convergence plots of optimization for the 100 generations experiment. In later iterations, the gap between the fitness and the model scores is closed, which implies the regularization term is reduced. Note that the raw model scores on average still converge, implying that the divergence for 20 and 100 iterations is also explained by stochasticity across runs.

As the optimizer converges, performance becomes constrained by the regularization penalty and the diminishing exploration variance of CMA-ES. Once a sufficient local region has been found, the optimizer reduces its step size and focuses on refining that region. Any marginal log-probability gained from exploring secondary evidence may be outweighed by the penalty of expanding the mask. On top of that, because insertion and deletion are optimized separately and their solutions may also focus on different regions, combining them through intersection can cancel found evidence. Consequently, a larger search budget merely seems to allow the optimizer to better exploit a region of interest, leaving AUC metrics stable but capturing only marginal extra evidence.

Hierarchical Optimization & Pre-softmax logits. As Table 5 shows, using a hierarchical optimization strategy is beneficial to final performance, as noted by especially better deletion scores (drops from .593 to .494). This confirms the intuitive benefit behind hierarchical optimization, as it should prevent the CMA-ES search distribution to collapse onto more fine-grained noise in the early generations and promote the mean vector $m^{(g)}$ to converge towards meaningful regions. The effect of pre- vs. post-softmax logits inside our objective is more nuanced compared to hierarchical optimization, yet still we observe a boost in performance (.721/.494 vs. .700/.525 for insertion and deletion). Pre-softmax logits mathematically better reflect the internal decision process inside the model, as the softmax function may squash the raw logits into a probability range.

Table 5: Ablation study results of hierarchical optimization and using pre-softmax logits on CLEVRER. Both hierarchical optimization and using pre-softmax are beneficial to final performance.† denotes the default configuration shared across ablation groups.

Hierarchical Ablation

Parameter	Value	Sufficiency ↑	Necessity ↑	AUC Ins ↑	AUC Del ↓
Hierarchical	True†	**.695** ± .299	**.430** ± .368	**.721** ± .212	**.494** ± .219
Hierarchical	False	.640 ± .296	.415 ± .351	.642 ± .296	.593 ± .224

Pre-Softmax Ablation

Parameter	Value	Sufficiency ↑	Necessity ↑	AUC Ins ↑	AUC Del ↓
Softmax	Pre†	**.695** ± .299	.430 ± .368	**.721** ± .212	**.494** ± .219
Softmax	Post	.653 ± .292	**.427** ± .354	.700 ± .201	.525 ± .221

Table 6: Metrics for optimization and clustering strategies for the **Simple** dataset. Results are grouped by optimization objective to highlight the effect of Spatial (2D) vs. Spatiotemporal (3D) clustering. Overall, 2D Spatial clustering consistently promotes sufficiency and insertion, while spatiotemporal clustering provides better necessity and deletion scores across both objectives.

Objective	Clustering	Sufficiency ↑	Necessity ↑	AUC Ins ↑	AUC Del ↓
Joint	2D Spatial	**.966** ± .087	.441 ± .350	**.925** ± .158	.459 ± .257
	3D Spatio-temp	.954 ± .174	**.531** ± .352	.890 ± .167	**.403** ± .294
Separate	2D Spatial	**.859** ± .225	.291 ± .285	**.922** ± .136	.455 ± .274
	3D Spatio-temp	.827 ± .293	**.331** ± .279	.890 ± .176	**.354** ± .253

Spatio-Temporal Clustering and Mask Formation. On the *Simple* dataset (Table 6), spatial 2D clustering consistently outperforms 3D clustering in terms of insertion and sufficiency for both the separate and joint objective, by achieving near-perfect sufficiency (.966 for joint, .859 for sep.) and higher insertion AUC (.925 for joint, .922 for sep.) Meanwhile, both deletion AUC and necessity fall behind compared to spatiotemporal clustering, indicating the spatial clustering falls short in capturing *all* evidence. For *Simple*, a few 3D clusters capture the salient object across time, as illustrated in Figure 4. This provides a simple optimization target that when hidden completely destroys evidence, aligning with the better necessity and deletion metrics we find for the 3D clustering strategy. When using a 2D strategy, different spatial clusters distributed over frames individually contain evidence which in turn provide easy targets for insertion, yet complicate optimization for deletion as it introduces redundancies across time.

Moving from *Simple* to *CLEVRER* (see Table 7) introduces increased optimization complexity due to multi-object interactions and more dynamic scenes. Here, under the *joint* optimization objective the metrics for spatiotemporal clustering decrease significantly compared to only spatial clustering. As CLEVRER contains more dynamic

Table 7: Metrics for optimization and clustering strategies for the **CLEVRER** dataset. Results are grouped by optimization objective to highlight the effect of Spatial (2D) vs. Spatiotemporal (3D) clustering. Under the joint objective, performance of spatiotemporal clustering is strongly decreased, while this approach becomes most robust on average for the separate objective. Joint optimization seems to benefit from 2D flexibility in dynamic scenes, while separate optimization needs 3D coherence to avoid cancellation by intersection.

Objective	Clustering	Sufficiency ↑	Necessity ↑	AUC Ins ↑	AUC Del ↓
Joint	2D Spatial	**.887** ± .142	**.478** ± .346	**.785** ± .137	.662 ± .218
	3D Spatio-temp	.725 ± .338	.418 ± .343	.635 ± .212	**.592** ± .293
Separate	2D Spatial	.633 ± .334	.411 ± .354	**.823** ± .142	.619 ± .186
	3D Spatio-temp†	**.695** ± .299	**.430** ± .368	.721 ± .212	**.494** ± .219

multi-object scenes, 3D SLIC may group together evidence that should remain separate or produce tubelets that are too noisy for the underlying reasoning task (see Figure 4). Frame-wise clustering may be better suited to CLEVRER in such cases, because it allows the optimizer to isolate localized evidence without forcing it into a temporally coherent tubelet.

However, this trend reverses again when applying the *separate* optimization strategy as now 3D spatiotemporal clustering becomes the most robust approach. Under the separate objective, individual separate optimization runs may focus on different subsets of the input. Because the final mask is now constructed by intersection, evidence from each independent run may cancel out. This effect is especially evident in frame-wise clustered settings (*2D Spatial*) where each run is allowed to focus on completely different frames. Consequently, spatiotemporal clustering interacts with the separate objective in a way that improves performance as it prohibits scattered masks.

Overall, these results indicate that hierarchical optimization, a pre-softmax objective and spatiotemporal clustering combined improve optimization performance. However, none of the tested configurations fully solve the objective. Expanding the mask around localized evidence becomes problematic due to a complex optimization landscape and the regularization terms that induce a concise mask.

5.2 Solution Consistency

Subsequently, we are interested whether optimal solutions with respect to the perturbation objective reflect real evidence for the LVLM. Ideally, optimal solutions consist of highly similar masks. If this is not the case, it indicates that multiple explanations can equally account for the model's behavior, which would be problematic as it becomes unclear what an explanation truly captures and what regions actually account for the model's output.

Within Table 8, we show consistency of solutions across runs and within populations. In the first part of the table, the population dynamics within the final generation are shown. Interestingly, we observe that with a population size of 20, up to 10 candidates

Table 8: Consistency and performance metrics. **Population Dynamics** measure the similarity of candidate solutions within the given fitness margin during a single run (pop. size = 20). **Inter-Seed Consistency** evaluates the stability of the final masks and metrics (Mean ± Std) across 10 random seeds. Clearly, the consistency across runs is extremely limited. With $N = 120$, the top 2% of tubelets consists of 2-3 items, of which on average around half of these are shared across random seeds. This fraction decreases even further when analysing the top 10%. Combined with the fact that, especially for 'Simple' and CV-Bench, the objective scores are often highly similar, this is a strong indication of ill-posedness of the problem setup. Multiple substantially different masks can achieve very similar scores with respect to the perturbation objective.

	Population Dynamics			**Inter-Seed Consistency (10 Runs)**			
Dataset	Top Cand. ↓	IoU@10% ↑	IoU@2% ↑	IoU@10% ↑	IoU@2% ↑	**Sufficiency** ↑	**Necessity** ↑
CV-Bench	7.6	.544	.395	.128	.411	.748 ± .060	.698 ± .005
Simple	9.8	.384	.408	.302	.352	.809 ± .106	.312 ± .058
CLEVRER	7.8	.527	.572	.219	.170	.627 ± .288	.414 ± .301

represent a solution which is within 10% fitness range of the best found solution, although this is dependent on the data. More importantly, within this set of similar solutions, we find IoU values of at most 0.544 for the top 10% of tubelets and .408 for the top 2% of tubelets. After derivation, we find that $IoU = \frac{x}{2-x} \approx 0.5$ provides $x \approx 0.7$ so the number of shared tubelets across the population is at most roughly 70%. In practice, this suggests that while the optimizer consistently identifies a core set of evidence (e.g., a primary salient object), it arbitrarily includes different surrounding regions across candidates to achieve similar fitness scores. CMA-ES finds multiple heatmaps that each contain a similar amount of evidence, while these heatmaps are not necessarily aligned.

The right part of the table shows solution consistency when formulating heatmaps over different random seeds and with different starting points for each run. Dataset *Simple*, with very local evidence, shows roughly 50% of shared tubelets over solutions for the top 2%. For the broader context (top 10%), IoU scores drop towards .288 and approximately 40% of tubelets remain shared. This share lowers for the more complex CLEVRER dataset and the static CV-Bench image data. However, the higher standard deviations show the method largely finds different heatmaps with different quality.

These low intersection scores suggest the objective landscape is multimodal and lacks a unique minimum. Consequently, the model's predictive confidence can be equally broken or restored by entirely disjoint sets of visual evidence, making it impractical to pinpoint one true causal region.

We wonder whether this set of results is limited to our method. Therefore, we also apply iGOS on CV-Bench image data following Xing et al.'s original benchmark data [56]. Results are shown in Table 9 and Figure 6. Observe that consistency across runs is extremely limited with IoU scores of .199 for the top 10% salient pixels and only .068 for the top 2%. While on average, the AUC scores fluctuate with wider standard deviations, on data-point level we observe cases that show non-uniqueness of optimal masks. For instance, the labeled '0' data-point in the CV-Bench dataset is, after 10 independent runs, characterized by Deletion AUC of 0.334 ± 0.091 and Insertion AUC

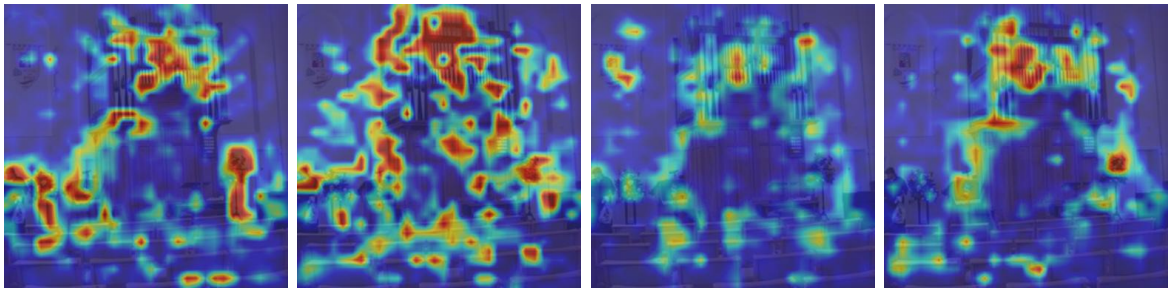


Figure 6: Four independent runs of iGOS [56] on the same image (CV-Bench, "How many organs are in this image?"). With visual inspection, roughly the same set of regions is identified, with different regions around it. However, the IoU of the top 2% and top 10% regions is respectively 0.041 and 0.149. More interestingly, the standard deviation around the deletion and insertion AUC is only 0.091 and 0.031 respectively; Each heatmap has a highly similar quality score while differing significantly in identified crucial regions.

Table 9: *iGOS consistency*. Apart from our own method, we also investigate uniqueness of solutions with iGOS. Again, the low IoU shares indicate convergence towards distinct solutions across runs. For * we use an extended version of iGOS for video input, while for CV-Bench, we measure on the untouched, original implementation.

Dataset	IoU@10% ↑	IoU@2% ↑	AUC Ins ↑	AUC Del ↓
CV-Bench	.199	.068	.925 ± .175	.334 ± .091
Simple*	.159	.352	.880 ± .067	.579 ± .190
CLEVRER*	.185	.074	.775 ± .210	.490 ± .328

of 0.778 ± 0.031 while Top 2% IoU equals 0.041 and Top 10% IoU 0.149. In other words, each solution independently approaches optimal values as identified by the good AUC metrics and very low standard deviations, while the most salient set of pixels is almost disjoint across runs. We visualize this in Figure 6. The instability of solutions appears inherent to the perturbation objective and not to the chosen method.

Overall, we show across methods a high instability across random seeds, as mask converge towards diverging solutions with highly similar scores. As the perturbation objective allows multiple distinct near-optimal solutions, it does not guarantee a unified explanation for a video. Consequently, this implies that it becomes inherently unclear whether and when perturbation-derived heatmaps for LVLMs represent all true evidence.

5.3 Video Complexities

Next, we evaluate what real-world aspects of video question answering complicate the perturbation objective. We analyse disjoint aspects using CLEVRER classifications as proposed in Table 2. Results are shown in Table 10. Good necessity and deletion scores imply the mask hides a crucial set of local evidence to confuse the model, while good insertion and sufficiency imply the inverted mask inserts local evidence for output

Table 10: Quantitative comparison of distinct categories on CLEVRER. Each category requires an increased level of (temporal) reasoning, starting from static attribution. Note how the reasoning required by the category affects the necessity and sufficiency scores.

Category	**Sufficiency ↑**	**Necessity ↑**	**AUC Ins ↑**	**AUC Del ↓**	**% Correct**
A. Spatial Attribution	0.460	0.187	0.708	0.516	80.0%
B. Temporal Attribution	0.524	0.259	0.731	0.481	66.7%
C. Multi-object Interaction	0.491	0.334	0.706	0.404	93.3%
D. Causal Reasoning	0.559	0.345	0.662	0.427	70.0%
E. Predictive Reasoning	0.484	0.466	0.582	0.345	73.3%
F. Counterfactual Pert.	0.535	0.489	0.621	0.386	83.3%

recovery.

Categories **A** and **B** contain questions about the existence of stationary and moving objects. The high insertion values indicate the method locates a set of evidence that generally restores confidence quickly. However, the low sufficiency and extremely low necessity scores reveal the method struggles to capture a *complete* set of relevant visual evidence, although it performs slightly better when identifying moving objects (**B**).

For multi-object interactions (cat. **C**), necessity scores and deletion AUC improve substantially (i.e., from .516 to .404 for deletion). Intuitively, it is relatively easy to perturb, for instance, collisions so that the model no longer recognizes these, which could explain the better necessity and deletion scores. Consequently, within this setting, interactions appear relatively easy to faithfully capture as evidence.

When questions require more complex reasoning, such as finding the causal origin of collisions (cat. **D**), predicting future events (cat. **E**) or counterfactual reasoning (cat. **F**) this trend continues. Necessity scores increase up to .489 for cat. **F** and deletion scores improve up to .345 (for cat. **E**). Meanwhile, insertion scores drop (e.g., .582 for cat. **E**). For questions about future events, hypothetical scenarios or localization of causal evidence, removing small portions of evidence will have great effect on the final answer. For instance, when asked whether a cylinder will be hit by a ball, or what happens when a cube is removed, very local parts can destroy evidence when removed (e.g., the ball in a single frame). However, to construct the entire set of evidence to answer such questions, a lot of distributed context is required (e.g., the cylinder and the ball over multiple frames). Likely, this causes degraded insertion but improved necessity and deletion scores for such questions.

Within our setting, these results imply evidence is distributed in differing ways. We find that in many scenarios, optimizing a heatmap to be either sufficient or necessary requires significantly more context than its counterpart. However, the perturbation formulation explicitly expects evidence to adhere to both simultaneously. When similarly penalizing a mask to be concise, either one objective will be prioritized during optimization. This mismatch in required real-world evidence complicates the convergence towards a faithful mask that captures a complete set of evidence. This result may explain why masks diverge across runs, as the optimizer may favor either insertion or deletion arbitrarily.

Table 11: Performance of our optimal configuration (†) against baselines across datasets. The random baseline does not explicitly optimize for sufficiency and necessity, while iGOS shares the same objective as our method, but adapts a different optimization methodology. Observe that for *deletion*, our method outperforms both baselines on Simple and CLEVRER. However, this is not the case for *insertion* and for Kinetics-400 our method performs roughly similar to the random baseline and worse to iGOS. This suggests our method works reasonably well in simple videos where it finds the most important parts to break the model's confidence when hidden (better deletion AUC), yet in complex videos like Kinetics-400, it breaks down. Likely, evidence is too distributed to find a suited solution using Evolutionary Strategies. For *, we use a manually extended version of iGOS for video input. The bottom half, denoted with $^+$, includes results from SWAG-V [17] as *reference*, as they experiment using **3D CNNs**.

	Simple		**CLEVRER**		**K400**	
Method	**AUC Ins ↑**	**AUC Del ↓**	**AUC Ins ↑**	**AUC Del ↓**	**AUC Ins ↑**	**AUC Del ↓**
Random	.700	.775	.617	.587	.598	.593
iGOS*	**.880**	.579	**.758**	.580	**.662**	**.471**
Ours	.827	**.354**	.721	**.494**	.616	.603
SWAG-V$^+$	-	-	-	-	.262$^+$	.068$^+$
G-CAM$^+$	-	-	-	-	.315$^+$	.136$^+$

5.4 Comparison

The subsequent set of experiments compares our method against a randomized baseline and iGOS [56] adapted to video. Results across three datasets are summarized in Table 11.

Quantitative Results. Interestingly, our method does not consistently outperform iGOS metric-wise. On both Simple and CLEVRER, our method outperforms iGOS only in terms of *deletion* and is worse in terms of *insertion*. This means ours is better able to localize a crucial input set that successfully destroys evidence when hidden, while it is incomplete on its own. As both methods differ in optimization regime rather than objective, the gradient-based search of iGOS seems better at finding the distributed context needed to rebuild model confidence.

This trend continues for the K400 dataset, where iGOS outperforms our method and we approach randomized values for both insertion and deletion. Both CLEVRER and K400 increasingly require evidence that is distributed across multiple clusters and frames. Likely, as our results for zero regularization and more iterations show, our method struggles to find a complete set of distributed evidence over multiple tubelets. This effect is especially evident within K400, as our method completely breaks down. Our method only provides localized approximations of evidence for data of moderate complexity.

Interestingly, compared to methods adapted for 3D CNN's, deletion scores are rather high for both our method and iGOS. However, none of these methods are tested across both LVLMs and 3D-CNNs for which this divergence can be attributed to both the method itself and the LVLM.

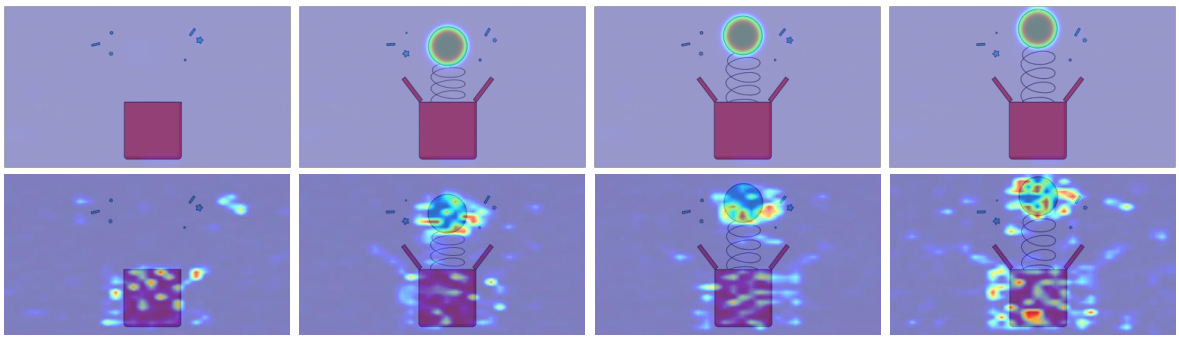


Figure 7: Exemplary qualitative comparison of our method (top) against that of Xing et al. [56] (iGOS) on dataset Simple. The attached questions asks: "What color does the central moving object have"? Insertion / Deletion AUC metrics are highly similar: .951/.362 (Ours) vs .910/.389 (iGOS). Observe how our construction using tubelets aligns with semantic concepts as shown by the exact highlight of the ball. However, iGOS suggests that this may not be the only concept contributing to the final explanation, as it also highlights the red box with similar scores.

Qualitative Results. Although quantitatively, our method does not differentiate much from iGOS, visual inspection reveals a significant difference in produced heatmaps. Figure 7 shows an example where, metric-wise, iGOS performs similar to our method. However, the iGOS heatmap captures distributed evidence across the video, while our method only localizes a single (salient) object. Note that both have highly similar quality scores and this further illustrates the instability of solutions under the perturbation objective.

More interestingly, in Figure 8, we provide an example heatmap for CLEVRER. In this scenario, our method produces an extremely localized heatmap, while iGOS provides a scattered heatmap failing to align with any actual evidence (underlined by the worse scores). In general, we find this trend consistent as iGOS heatmaps are scattered across videos and ours aligned with concepts over time, dependent on the complexity of the data. Additional examples are provided in Appendix D and suggest a similar trend. Interestingly, while clustering video input over tubelets is subjectively beneficial to interpretability, it does not inherently create more faithful heatmaps as underlined by the quantitative scores.

While we find that combining evolutionary search with spatiotemporal tubelets forces heatmaps into human-readable concepts, the gradient-based search of iGOS better captures the model's distributed context. Under the current perturbation objective, constraining for visual coherence in tubelets may risk leaving out scattered evidence the LVLM relies on. While CMA-ES is proposed to improve the optimization process, results indicate that our framework is not yet complete. Moreover, they point to an inherent bias in the explanations introduced by both the optimization strategy and the defined search space.

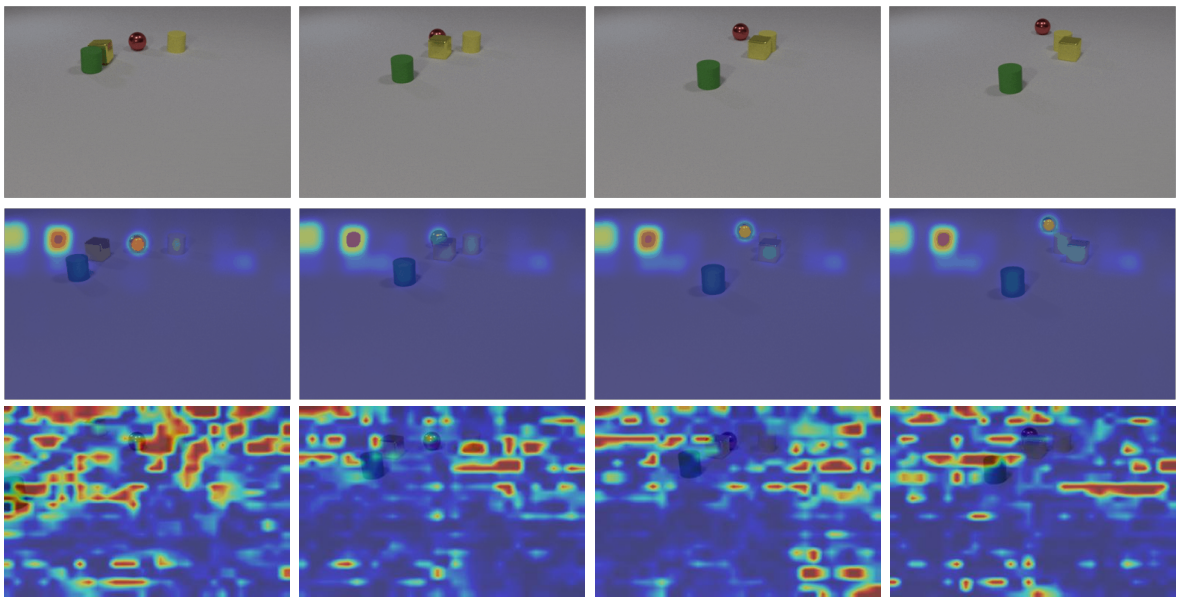


Figure 8: Additional exemplary comparison of our method (middle row) on CLEVRER against the method of Xing et al. [56] (bottom). Original video is shown in the top row with the question "Are there any moving green objects when the video begins"? In this example, our method (AUC Ins/Del: .924/.350) provides localized explanations whereas iGOS (.833/.749) fails in highlighting a concise set of evidence. Interestingly, both heatmaps do not highlight the exact green cylinder. The heatmap of our method only highlights a non-existent object and a moving red ball, suggesting the model bases itself on wrong evidence, as AUC scores suggest high quality.

6 Discussion

Across all ablations, none of the tested configurations satisfied the perturbation objective. Increasing optimization budget did not produce improvements and reducing the regularization term only improved sufficiency marginally. Similarly, when extending the state-of-the-art optimization-based perturbation method iGOS++ to videos, the method leads to results of only limited faithfulness for complex videos. Regardless of the chosen optimization regime, highlighting evidence for video question answering comes with notorious difficulties.

A key finding of our work is that we provide evidence that the perturbation objective allows for multiple near-optimal but disjoint solutions. Even when AUC metrics reflect near-optimal confidence drops, we find extremely low inter-seed consistency across independent runs (Table 8). This implies the optimization landscape contains multiple disjoint solutions that could each represent a high-quality mask metric-wise. Not only do these results doubt the existence of single unified explanation, they suggest that provided heatmaps will at most show an approximation of evidence for an LVLM's answer. This aligns with our theoretical analysis in Appendix A: mathematically, there is no guarantee that a unique global minimum exists for the objective function. There may be many partial masks $\hat{M} \neq \mathbf{1}$ that satisfy $f(\Phi(V, \tilde{V}, 1 - \hat{M}); Q) \leq f(\Phi(V, \tilde{V}, \mathbf{0}); Q)$. We cannot

attribute this solely to the choice of the evolutionary optimizer, as shown by the similar behaviour of iGOS to converge to non-unique masks (Figure 6). Problematically, it becomes disputable up to what part heatmaps over the video represent salient evidence.

Our analysis and experiments provide multiple explanations that hinder convergence towards meaningful unique solutions. First of all, the increased non-linearity and dimensionality due to video input and the highly non-linear function of the LVLM make optimization inherently more difficult. Subsequently, temporally redundant concepts increase the context dependence of saliency for input regions, which further complicates assigning saliency scores faithfully. Interestingly, while the usage of tubelets accounts for temporally redundant concepts, the combination with CMA-ES is not necessarily beneficial to the optimization regime. Ultimately, we find it hinders capturing distributed evidence as CMA-ES obtains regularization penalties that seem to prevent convergence and SLIC's 3D clusters may not be aligned with the model's evidence. A global search method like CMA-ES uncovers more disjoint local optima instead of offering a methodological advantage to outperform iGOS.

Our experiments on the CLEVRER dataset highlight that the individual necessity and sufficiency objectives do not always align and evidence is distributed in differing ways. Think of it like a Jenga tower: removing just a few parts can be destructive while you need a lot of parts to reconstruct the tower. The perturbation objective assumes (either explicitly or implicitly) a symmetric relationship between insertion and deletion of evidence, which does not hold anymore when evidence is context dependent or redundant. As a mask may favor either one of the objective (e.g., the part to destroy the Jenga tower vs. reconstructing the bottom half of it), convergence towards unique solutions becomes problematic as both parts contribute equally metric-wise. However, it is unclear from the heatmap alone what evidence is actually covered and what isn't.

As the perturbation objective allows non-unique solutions and the mechanisms of the LVLM for videos complicate the landscape for this objective, providing faithful heatmaps using perturbations becomes impractical when dimensionality increases. Ultimately, our findings suggest that relying on perturbation-based heatmaps for video LVLMs risks providing a false sense of interpretability, as multiple disjoint explanations can mathematically justify the same model behavior.

6.1 Limitations

An important limitation in our optimization setup is the final mask formulation for the separate objective using the Hadamard product. As shown, the separate optimization regime significantly reduced performance when combined with frame-wise clustering. In such cases, insertion and deletion masks may localize different subsets of evidence over different frames, for which the intersection by the Hadamard product will cancel out evidence. This may be an underlying issue throughout experiments. A possible improvement would be to jointly optimize both objectives while simultaneously promoting agreement on both masks, or concatenating masks by union.

We ran each experiment once over only 30 to 40 videos due to required compute access and high time complexity. Moreover, only three datasets and two models are evaluate, which combined limit generalization of our results. Time complexity is another

important limitation, as runs take up to multiple minutes to produce a single faithful heatmap, making our method less useful in practice. Moreover, the high std. scores indicate the set of results do not provide a complete picture, for which we should include more runs over a larger number of videos. Similarly, some settings limit the generalization of the results: the decision to stick with SLIC clustering, the un-ablated number of tubelets, or the decision of the CMA-ES optimizer for which recent work provides improvements (e.g., BI-POP [33]). However, we must note that even with these limitations, underlying issues for the perturbation objective remain.

6.2 Future Directions

Future work should focus on addressing convergence issues, for instance by including a dynamic regularization term or, for example, BI-POP, to prevent over-exploitation. Most importantly, it will be interesting to focus on a different optimization methodology (i.e., different objective, final mask formulation and optimizer), to capture both distributed and localized evidence. For this reason, SLIC tubelets could be replaced with clusters that are better aligned with the model's actual internal processing. This can be by perturbing semantic concepts inside the LVLM's cross-modal projector, within the latent space, to more faithfully align with the model's reasoning. This would simultaneously overcome the issue of temporal dependencies in pixel space, while it is harder to interpretably visualize.

However, to really overcome optimization-based perturbation issues, different actions may be required. For instance, instead of a derivative-free optimizer guessing regions, future frameworks could also use structural causal models (SCMs) over the video's object trajectories to formally test "do-interventions" (e.g., DO(REMOVE BLUE SPHERE)). While this alone does not necessarily solve the problems posed by the objective, it is possible to present these concepts in an interactive fashion. Earlier work has shown that this will significantly benefit to the final goal of both interpretability and faithfulness [4].

7 Conclusion

In this study, we extend perturbation-based explainable AI methods towards video question answering, by employing an evolutionary optimizer to discover the most salient perturbations to the model's output. When moving to video LVLMs, the perturbation-model output relationship is complicated by the temporal dependencies and the high non-linearity of the LVLM. Applying a hierarchical optimization strategy and optimizing over spatiotemporally clustered pixels eases this relationship, leading to heatmaps of improved quality. However, we find it is hard to expand the mask around localized parts of evidence. The minimization objective that is standard for perturbation methods allows for multiple non-overlapping solutions that are each optimal, while these do not necessarily reflect semantic relevance.

8 Acknowledgements

First of all, we would like to thank Dr. H.R. (Hazel) Doughty, who provided thoughtful but concise feedback along the process and drafts of this thesis. For experiments, this work was performed using the compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University, but also that of Snellius (Dutch cluster), for which we thank the University and LIACS specifically for allowing expensive compute budget.

Throughout this project, Generative AI models were used to complete code, propose sentence structures or suggest experiments. While useful, these tools are (currently) no stand-alone concepts and we found these often misinterpret, hallucinate or overcomplicate matters.

References

[1] Radhakrishna Achanta et al. "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2274–2282. DOI: 10.1109/TPAMI.2012.120.

[2] Saleh Afroogh et al. *Beyond Explainable AI (XAI): An Overdue Paradigm Shift and Post-XAI Research Directions*. 2026. arXiv: 2602.24176 [cs.CY]. URL: https://arxiv.org/abs/2602.24176.

[3] Jean-Baptiste Alayrac et al. "Flamingo: a visual language model for few-shot learning". In: *Advances in neural information processing systems* 35 (2022), pp. 23716–23736.

[4] Apoorva Aravindkumar et al. "Explainable AI in healthcare: a systematic review of XAI use cases in imaging, diagnostics, and rehabilitation". In: *Frontiers in Artificial Intelligence* 9 (2026), p. 1749527.

[5] Shuai Bai et al. "Qwen2. 5-VL Technical Report". In: *arXiv e-prints* (2025), arXiv–2502.

[6] Shuai Bai et al. "Qwen3-vl technical report". In: *arXiv preprint arXiv:2511.21631* (2025).

[7] Oren Barkan et al. "Visual explanations via iterated integrated attributions". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2073–2084.

[8] Deepshikha Bhati et al. "A survey on Post-Hoc Explanation Methods for XAI Visualization". In: *Authorea Preprints* (2025).

[9] Hila Chefer, Shir Gur, and Lior Wolf. "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 397–406.

[10] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[11] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. "Masked autoencoders as spatiotemporal learners". In: *Advances in neural information processing systems* 35 (2022), pp. 35946–35958.

[12] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. "Understanding deep networks via extremal perturbations and smooth masks". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 2950–2958.

[13] Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3429–3437.

[14] David Gunning et al. "DARPA's Explainable AI (XAI) program: A retrospective". In: *Authorea* 2021.1115 (2021). DOI: 10.22541/au.163699841.19031727/v1. eprint: https://www.authorea.com/doi/pdf/10.22541/au.163699841.19031727/v1. URL: https://www.authorea.com/doi/abs/10.22541/au.163699841.19031727/v1.

[15] Nikolaus Hansen. "The CMA Evolution Strategy: A Tutorial". In: (2005).

[16] Nikolaus Hansen and Andreas Ostermeier. "Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation". In: *Proceedings of IEEE international conference on evolutionary computation*. IEEE. 1996, pp. 312–317.

[17] Thomas Hartley et al. "Swag-v: explanations for video using superpixels weighted by average gradients". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 604–613.

[18] Thomas Hartley et al. "Swag: Superpixels weighted by average gradients for explanations of cnns". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 423–432.

[19] Sara Hooker et al. "A Benchmark for Interpretability Methods in Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

[20] Alon Jacovi and Yoav Goldberg. *Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?* 2020. arXiv: 2004.03685 [cs.CL]. URL: https://arxiv.org/abs/2004.03685.

[21] Ying Ji, Yu Wang, and Jien Kato. "Spatial-temporal concept based explanation of 3d convnets". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15444–15453.

[22] Felix Kares et al. "What makes for a good saliency map? Comparing strategies for evaluating saliency maps in explainable AI (XAI)". In: *arXiv preprint arXiv:2504.17023* (2025).

[23] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[24] Rémi Kazmierczak et al. "Explainability and vision foundation models: A survey". In: *Information Fusion* 122 (2025), p. 103184.

[25] Saeed Khorram, Tyler Lawson, and Li Fuxin. "iGOS++ integrated gradient optimized saliency by bilateral perturbations". In: *Proceedings of the Conference on Health, Inference, and Learning*. 2021, pp. 174–182.

[26] Satyapriya Krishna et al. "The disagreement problem in explainable machine learning: A practitioner's perspective". In: *arXiv preprint arXiv:2202.01602* (2022).

[27] Hao Li et al. "Visualizing the loss landscape of neural nets". In: *Advances in neural information processing systems* 31 (2018).

[28] Junnan Li et al. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.

[29] Xuhong Li et al. “M4: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities, and Models”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2023).

[30] Zhenqiang Li et al. “Towards visually explaining video understanding networks with perturbation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1120–1129.

[31] Bin Lin et al. “Video-llava: Learning united visual representation by alignment before projection”. In: *Proceedings of the 2024 conference on empirical methods in natural language processing*. 2024, pp. 5971–5984.

[32] Haotian Liu et al. “Visual instruction tuning”. In: *Advances in neural information processing systems* 36 (2023), pp. 34892–34916.

[33] Ilya Loshchilov, Marc Schoenauer, and Michèle Sebag. “Bi-population CMA-ES agorithms with surrogate models and line searches”. In: *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*. 2013, pp. 1177–1184.

[34] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI]. URL: https://arxiv.org/abs/1705.07874.

[35] OpenAI. *Hello GPT-4o*. Accessed: 2026-06-08. 2024. URL: https://openai.com/index/hello-gpt-4o.

[36] Yong-Hyun Park et al. “Geometric Remove-and-Retrain (GOAR): Coordinate-Invariant eXplainable AI Assessment”. In: *XAI in Action: Past, Present, and Future Applications*. 2023. URL: https://openreview.net/forum?id=gh69Bu7k48.

[37] Vitali Petsiuk, Abir Das, and Kate Saenko. “Rise: Randomized input sampling for explanation of black-box models”. In: *arXiv preprint arXiv:1806.07421* (2018).

[38] Forough Poursabzi-Sangdeh et al. “Manipulating and Measuring Model Interpretability”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445315. URL: https://doi.org/10.1145/3411764.3445315.

[39] Zhongang Qi, Saeed Khorram, and Fuxin Li. “Visualizing Deep Networks by Optimizing with Integrated Gradients.” In: *AAAI*. Vol. 34. 2020, pp. 11890–11898.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[41] Yao Rong et al. “Towards human-centered explainable ai: A survey of user studies for model explanations”. In: *IEEE transactions on pattern analysis and machine intelligence* 46.4 (2023), pp. 2104–2122.

[42] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[43] Matteo Salis et al. “Extending Model-Agnostic XAI Methods for Regression Tasks in Spatio-Temporal Domains”. In: *International Conference on Discovery Science*. Springer. 2025, pp. 301–315.

[44] Wojciech Samek et al. “Evaluating the visualization of what a deep neural network has learned”. In: *IEEE transactions on neural networks and learning systems* 28.11 (2016), pp. 2660–2673.

[45] Ramprasaath R Selvaraju et al. “Grad-CAM: visual explanations from deep networks via gradient-based localization”. In: *International journal of computer vision* 128.2 (2020), pp. 336–359.

[46] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps”. In: *arXiv preprint arXiv:1312.6034* (2013).

[47] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[48] Gemini Team et al. “Gemini: a family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805* (2023).

[49] Shengbang Tong et al. “Cambrian-1: A fully open, vision-centric exploration of multimodal llms”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 87310–87356.

[50] Zhan Tong et al. “Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training”. In: *Advances in neural information processing systems* 35 (2022), pp. 10078–10093.

[51] Tomoki Uchiyama et al. “Adaptive occlusion sensitivity analysis for visually explaining video recognition networks”. In: *arXiv preprint arXiv:2207.12859* (2022).

[52] Tomoki Uchiyama et al. “Visually explaining 3D-CNN predictions for video classification with an adaptive occlusion sensitivity analysis”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 1513–1522.

[53] Yongjie Wang et al. “Gradient based feature attribution in explainable ai: A technical review”. In: *arXiv preprint arXiv:2403.10415* (2024).

[54] Zerui Wang and Yan Liu. “STAA: Spatio-Temporal Attention Attribution for Real-Time Interpreting Transformer-Based AI Video Models”. In: *IEEE Access* (2025).

[55] Rosina O Weber et al. “XAI is in trouble”. In: *AI Magazine* 45.3 (2024), pp. 300–316.

[56] Xiaoying Xing et al. “Where do Large Vision-Language Models Look at when Answering Questions?” In: *arXiv preprint arXiv:2503.13891* (2025).

[57] Kexin Yi et al. “CLEVRER: CoLlision Events for Video REpresentation and Reasoning”. In: *International Conference on Learning Representations*. 2020.

[58] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[59] Jianming Zhang et al. “Top-down neural attention by excitation backprop”. In: *International Journal of Computer Vision* 126.10 (2018), pp. 1084–1102.

[60] Xu Zheng et al. “F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI”. In: *ICLR 2025* (2025).

A Theoretical Analysis

We analyse the function $f(\Phi(V, \tilde{V}, M); Q)$, which maps mask perturbations to the model's log-likelihood score. Ideally, changes in M would correspond to smooth and consistent changes in f, reflecting the semantic relevance of regions (the genuine importance of features). In practice, this formulation presents two challenges: the mathematical complexity of the optimization landscape and a gap between the objective and true semantic relevance.

Non-uniqueness From a well-posedness perspective, the proposed optimization problem does not guarantee uniqueness or stability. While $M = \mathbf{0}$ (removing all visual evidence) is a natural candidate for minimizing this score, there is no theoretical guarantee that this is the (only) optimum. There may exist (multiple) $\hat{M} \neq \mathbf{0}$ such that

$$f(\Phi(V, \tilde{V}, \hat{M}); Q) \leq f(\Phi(V, \tilde{V}, \mathbf{0}); Q),$$

indicating that an partial set of perturbations may actually provide a better score than removing all evidence.

Complexity Subsequently we investigate the smoothness of the optimization landscape, for which multiple factors introduce complexities. *(1) Context dependence:* self-attention combines input tokens over the entire sequence and effectively couples all features together. As a result, the objective f is non-separable with respect to M: the marginal effect of a single element $M_{x,y,z}$ depends entirely on the surrounding mask. Consequently, local gradients are context-dependent, and small changes in M can lead to different optimization directions under different mask configurations. This effect is even more evident in videos, where the context of a feature in one frame is determined by all preceding and following frames. *(2) Non-linearity:* while context-dependence alone does not lead to non-convex optimization landscapes, the (masked) input is processed by a

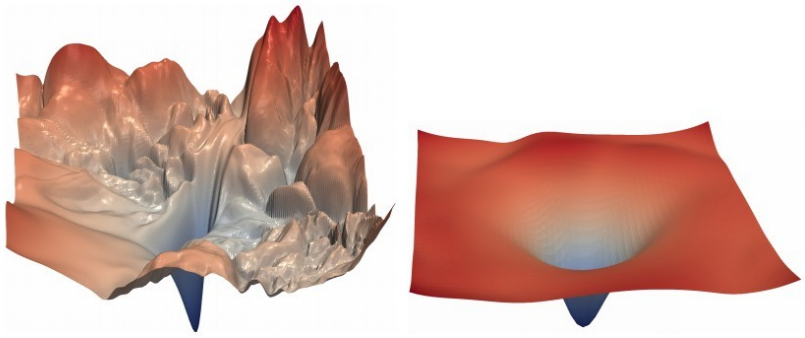


Figure 9: Two illustrative loss landscapes. On the right, the gradients points consistently to the global minimum. The figure on the left shows a more likely scenario for the formulation under Equation 4, where the function is non-smooth and gradients can be misleading towards (many) local minima. In reality, without clustering, the number of dimensions increases with the input size of the video (number of frames and frame size), making optimization extremely complex. Images from [27].

large combination of non-linear operations (e.g., attention softmax and feed-forward activations), which *does* promote irregularities in the optimization landscape. While this is not unique to LVLMs, the impact increases with the number of interacting and semantically relevant regions, as well as with stronger spatial-temporal dependencies. *(3) High dimensionality:* moving from images to video introduces an extra temporal dimension, which further increases the effective dimensionality of M. With higher dimensionality, the optimization difficulty increases. Combined, when moving from image towards video masking, small local changes in M can increasingly lead to non-smooth changes in the higher-dimensional f (Figure 9), which significantly increases optimization difficulty.

Semantic Relevance Even if we could reliably locate the global minimum M^*, it is not guaranteed to represent the genuine real-world importance of the visual regions. This disconnection originates from several aspects. *(1) Out-of-distribution inputs:* the interpolation $\Phi(V, \tilde{V}, M)$ can produce visuals that differ strongly from the training distribution [19, 36]. This may drop the model's certainty largely due to network confusion instead of the removal of actual evidence. However, the setup does not separate between these two, which weakens the link between mask values and semantic relevance. *(2) Linguistic priors:* many output tokens rely heavily on language priors instead of actual visual evidence [56]. In these cases, gradients with respect to the mask may diminish ($\|\nabla_M f\| \approx 0$), creating flat regions where the optimal mask value becomes arbitrary. *(3) Regularization terms:* our regularization terms ($\|M\|_1$ and $\mathrm{TV}(M)$) explicitly promote connected and localized evidence. While convenient for interpretability, this does not necessarily align with the nature of real-world visual concepts in videos. The regularization terms therefore introduce another disconnection with actual semantic relevance. *(4) Softmax squashing:* while the impact may be limited, it is worth mentioning that optimizing towards the (post-softmax) LLR does not truly reflect the internals of the model. In a toy example, output logits of [0, 15] and [0, 30] both yield a probability near 1.0 after converted through the softmax function. When

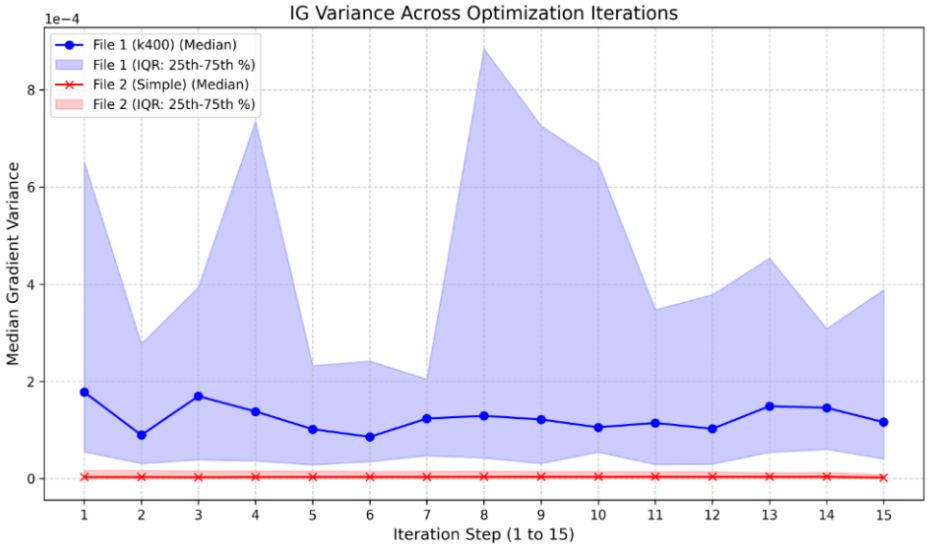


Figure 10: The variance of gradients along the path analysed over 20 samples of the *K400* (blue) and *Simple* datasets (in red), among different optimization iterations of iGOS++ [56]. Observe how the gradient is order(s) of the magnitude larger for k400, indicating a significantly more complicated loss landscape.

gradually hiding evidence to the model, true logits may drop significantly while the actual objective may remain similar.

Consequently, due to the increased dimensionality and temporal dependencies within videos, our objective becomes increasingly complex to optimize. Even if found, there is no guarantee that the mathematical minimum serves as true reflection of semantic relevance, as this may not be a unique solution, can be affected by OOD artifacts and influenced by the regularization terms. It should therefore be considered as approximation.

By applying the method only to visually grounded answers and similarly adjusting the objective to the set of visually relevant tokens within answers, we improve the robustness of our method. We reduce the dimensionality and context-dependence of M via spatiotemporal clustering and the hierarchical clustering prevents becoming stuck in local minima early on. Moreover, by using CMA-ES as a derivative-free optimizer we are less sensitive to uninformative gradients.

B Complete Experimental Setup

A full overview of additional parameters that are consistently used throughout experiments can be found in Table 12. All experiments are performed either via Snellius (the Dutch national supercomputer) or ALICE (Leiden University compute cluster), using NVIDIA A100 GPUs. For reproducibility, we set the random seed to '40', although this only has effect when applying this seed within the CMA-ES declaration. This is currently omitted because of the consistency experiments.

The number of tubelets (120) is based on findings in SWAG-V [17]. It is common for the population size of CMA-ES to take around $\frac{1}{5}$ of the dimension (≈ 20) [15]. The

LLR treshold, probability drop ratio and short sequence limit are directly obtained from Xing et al. [56].

Table 12: Complete Experimental Setup. Note: For the Keyword Selection strategy (LLR of 1.0), a word is only selected as a keyword if the model is at least ~ 2.72 times more confident in predicting it when looking at the original video compared to when looking at the blurred baseline.

Parameter	**Value**	**Description**
Spatio-Temporal Search Space (SLIC)		
Sampled Frames	8	Total number of frames extracted per video.
Number of Segments	120	Number of tubelets ($\mathcal{T}$) per video.
Compactness	10.0	Controls the shape of the superpixels.
CMA-ES Parameters		
Super Clusters	20	Number of super clusters for the first hierarchical stage.
Population Size (λ)	20	Number of candidate masks generated per iteration.
Initial Step Size (σ_{init})	1.0	Initial global step size for the CMA-ES search distribution.
Search Bounds	$[-5.0, 5.0]$	CMA-ES constraints for candidate weights before sigmoid.
Init. Weights (Insertion/Deletion)	-2.0 / 2.0	Starting values for the respective optimization modes.
Keyword Selection Strategy Parameters		
LLR Threshold (α)	1.0	Minimum log-likelihood ratio for a token to be visually grounded.
Short Sequence Limit	≤ 4 tokens	For short sequences, all valid tokens are evaluated without LLR filtering.
Evaluation and Masking Operation Parameters		
Baseline Type ($\tilde{V}$)	Blur	Uninformative baseline used to replace obscured visual evidence.
Prob. Drop Threshold	0.30	Drops videos where the baseline fails to reduce confidence by this margin.

C Full Hyperparameter Study

In Table 13, we showcase each run for the relevant set of hyperparameters for completeness. For the initial global step size, using 1.0 is most beneficial, while all other ablations are mostly discussed during the results section.

Table 13: Ablation study on CLEVRER. † denotes the default configuration shared across ablation groups. Notice how a stronger regularization limits the effect of especially the sufficiency within the optimization, while it strongly benefits the AUC metrics. Stronger regularization promotes semantic relevance. Interestingly, more iterations leads to worse performance on especially sufficiency; which suggests the optimizer overfits towards local optima ruled by regularizations inside the loss landscape.

Parameter	**Value**	**Sufficiency ↑**	**Necessity ↑**	**AUC Ins ↑**	**AUC Del ↓**
λ_{reg}	0	.729 ± .316	.432 ± .361	.585 ± .223	.636 ± .183
λ_{reg}	1	**.737** ± .304	.413 ± .347	.695 ± .198	.580 ± .225
λ_{reg}	5	.689 ± .292	.395 ± .354	.721 ± .200	.497 ± .230
λ_{reg}	10†	.695 ± .299	.430 ± .368	.721 ± .212	.494 ± .219
λ_{reg}	20	.642 ± .318	**.449** ± .351	**.733** ± .196	**.478** ± .246
Iterations	20	**.700** ± .197	.431 ± .348	**.721** ± .172	.519 ± .197
Iterations	50†	.695 ± .299	.430 ± .368	**.721** ± .212	**.494** ± .219
Iterations	100	.589 ± .338	**.433** ± .361	.699 ± .189	.496 ± .242
$\sigma^{(0)}$	0.5	.670 ± .337	.429 ± .357	**.732** ± .215	**.470** ± .216
$\sigma^{(0)}$	1.0†	**.695** ± .299	**.430** ± .368	.721 ± .212	.494 ± .219
$\sigma^{(0)}$	2.0	.694 ± .276	.419 ± .350	.676 ± .218	.554 ± .223
Hierarchical	True†	**.695** ± .299	**.430** ± .368	**.721** ± .212	**.494** ± .219
Hierarchical	False	.640 ± .296	.415 ± .351	.642 ± .296	.593 ± .224
Softmax	Post †	**.695** ± .299	.430 ± .368	**.721** ± .212	**.494** ± .219
Softmax	Pre	.653 ± .292	**.427** ± .354	.700 ± .201	.525 ± .221

D Additional Qualitative Results

To provide a complete and honest insight in the results of our method, we provide additional qualitative results in Figures 11 and 12. Notice that our method is often able to highlight localized evidence, while this is often not the complete set of evidence over the entire video. When complexity increases (K400), the exact alignment with semantic relevance decreases, although this is a subjective measurement.

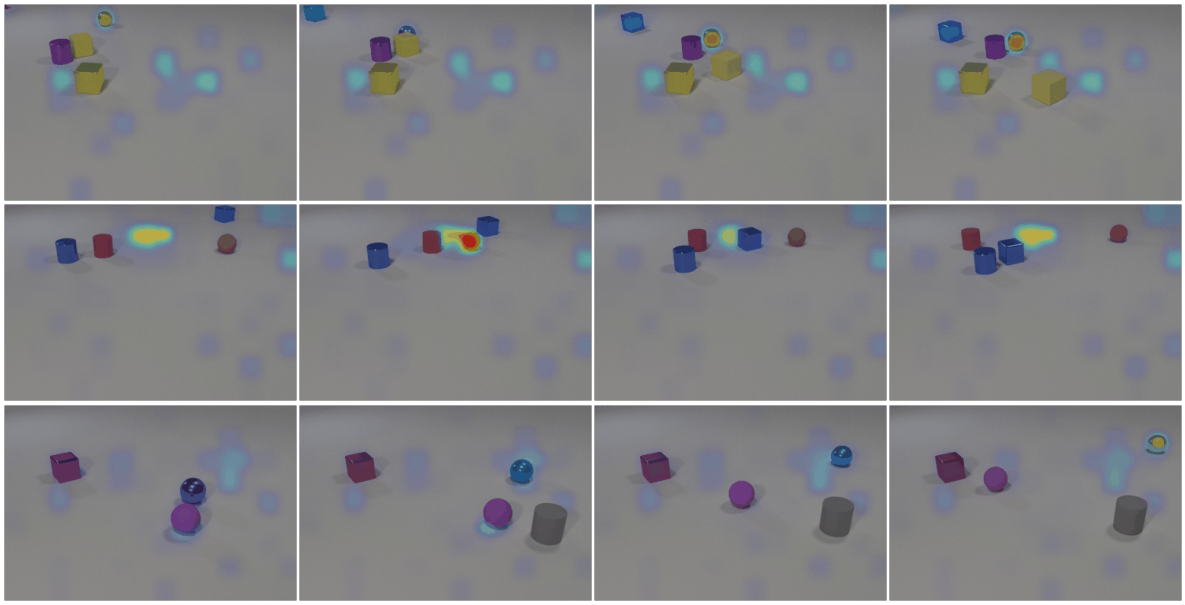


Figure 11: Example heatmaps provided by our method on CLEVRER data. From top to bottom, corresponding questions are: "Are there any stationary cubes when the cylinder enters the scene?" ("Yes, one yellow cube", ground truth: yes, AUC Ins/Del: 0.795/0.677). Second row: "Are there any moving rubber objects when the video begins?" ("Yes, yes", ground truth: no, 0.890 / 0.440. The model gains quick confidence by inserting the highlighted region that is only showing a moving object at the second frame. Third row: "Are there any moving rubber objects?" ("Yes, two.", ground truth: yes, 0.747/0.504). Interestingly, the heatmap highlights a non-rubber object.

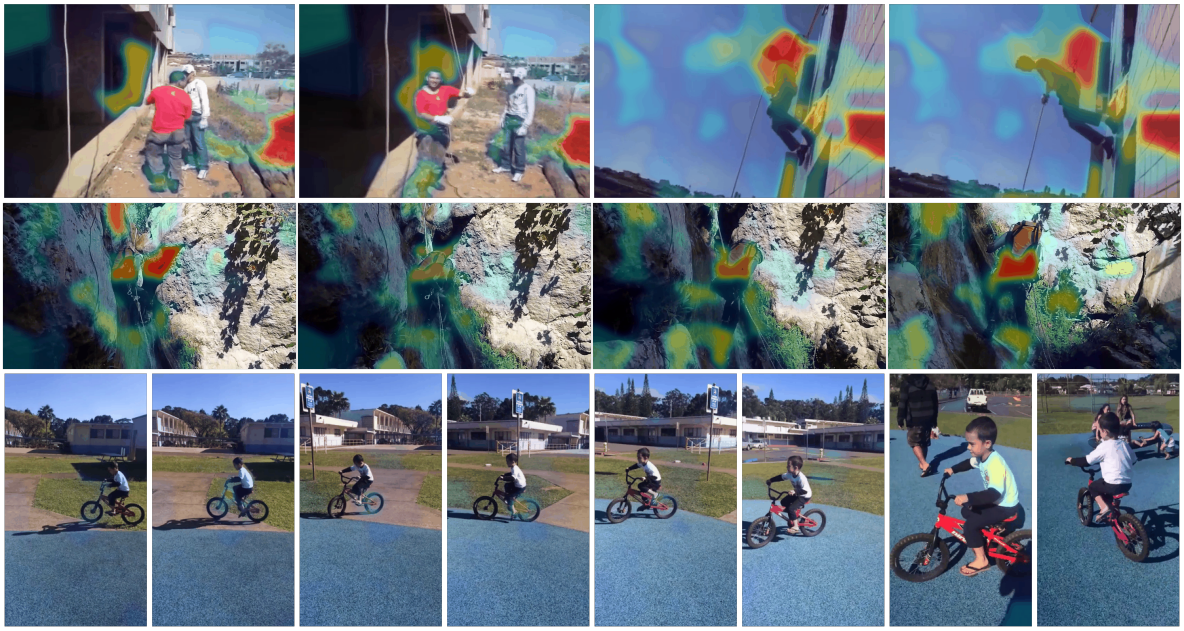


Figure 12: Heatmaps after running our method on the K400 dataset. Notice how the method is able to identify localized queues (the climber, abseiler and wheels of the bike). However, these queues do not reflect a complete set of evidence on their own.