

A. Computational Analysis

To substantiate our claim of a "modest computational footprint" and provide full transparency, this section details the computational costs of our models. All models were trained on a single NVIDIA-A40 GPU.

Table 1 provides a breakdown of parameters and computational load (GFLOPs) for our minimal pair. The addition of our temporal STI module adds approximately 50.5M trainable parameters and 88.6 GFLOPs over the spatial-only baseline. Both models leverage the same large, frozen 303.5M parameter backbone, ensuring a controlled comparison.

Table 2 contextualizes our models' computational costs against other recent state-of-the-art video saliency models. Despite using a large backbone, our models are computationally efficient, particularly our spatial-only variant, which has the lowest GFLOPs among the compared methods.

B. Expanded Methods

B.1. Experimental Design Rationale

Our controlled experiment was designed to unambiguously isolate and quantify the contribution of temporal information in video saliency prediction. We used well-established contemporary architectures to ensure that our findings generalize to modern approaches. The core principle is a "clean-room" experimental design: by keeping the powerful backbone, readout and finalizer heads, and training procedures constant across both models (UniFormerSal-ST and UniFormerSal-S), any performance difference can be directly attributed to the single variable of interest—the integration of temporal cues. This approach allows for causal analysis of when and why temporal information helps or hurts video saliency prediction, moving beyond correlational observations that confound architectural differences with temporal processing effects.

B.2. Model Architecture Details

This section details UniFormerSal-ST and its spatial-only ablation, UniFormerSal-S. The STI module is our mechanism for temporal fusion and represents the key difference between our ST and S models. It addresses the core challenge of injecting the global temporal context from the single class token into the detailed, per-frame spatial maps without degrading spatial resolution. We chose cross-attention over simpler fusion methods (element-wise addition, concatenation followed by convolution) because it enables dynamic, context-aware integration. The cross-attention mechanism allows spatial feature map tokens to actively query a condensed representation of global temporal information, dynamically weighting the importance of different aspects of the global temporal context for each spatial location. Spatial feature map tokens, derived from ViT layers 20-23 via a Dynamic

Position Encoding block, act as Queries, attending to the global spatio-temporal token, which is linearly projected to provide the Keys and Values. This operation allows every spatial location to "pull" the most relevant temporal information from the global summary, modulating its features accordingly. This directed fusion is crucial for adapting general video features to the nuanced task of saliency prediction while preserving the computational efficiency of the frozen backbone.

B.2.1. UniFormerV2 Architecture and Feature Decoupling

UniFormerV2's architectural philosophy naturally produces two distinct feature streams: high-resolution, per-frame spatial maps (local information) and a single, condensed class token that summarizes the video's global dynamics. The architecture's modular design is key to our controlled experiment, as it preserves the core spatial feature extractors of a pre-trained image ViT while cleanly separating temporal processing into two distinct modules: a local temporal multi-head relational aggregator (MHRA), referred to as Local UniBlockV2, for frame-to-frame temporal dynamics and a global spatio-temporal cross MHRA, referred to as Global UniBlockV2, that distills a video-wide summary class token. This architectural decoupling of information into detailed, per-frame spatial maps and a compact, global summary token presents both an opportunity and a challenge. Naively combining these streams can dilute critical spatial detail, which motivated our cross-attention-based STI design. Our work introduces this powerful pattern to the video saliency domain as a targeted solution for integrating global temporal context with fine-grained spatial information.

B.3. Training Details

The predictor networks for UniFormerSal-ST, UniFormerSal-S, and other ablation variants were trained exclusively on the LEDOV dataset. We used input clips of 16 frames (30fps, resized to 336x336) and trained for 20 epochs with a batch size of 4. Following recommendations by [9] for Adafactor, we utilized the first momentum to compute the exponential moving averages of gradients and did not scale the learning rate by the L2-norm of the weights. The initial learning rates for the trainable components were set as follows: $2e-4$ for the STI module (if active), $2e-3$ for the readout network, and $2e-3$ for the finalizer stage. No learning rate decay schedule was applied. We trained on a single NVIDIA-A40 GPU. The Gaussian smoothing applied in the finalizer stage utilized a learnable sigma, initialized to [16.0, 16.0].

B.4. Evaluation Metrics and Baseline Definitions

This section provides detailed explanations of the evaluation metrics and baseline models that were compressed in the main paper due to space constraints.

Table 1. Computational footprint of our minimal pair models. The addition of the STI module for temporal processing represents a modest increase in trainable parameters and GFLOPs.

Model	Trainable Params	Non-Trainable Params	Total GFLOPs
UniformerSal-S	89.2 M	303.5 M	122.2
UniformerSal-ST	139.7 M	303.5 M	210.8

Table 2. Comparison of total parameters and GFLOPs against other SOTA video saliency models. Our models demonstrate competitive computational efficiency.

Model	Total Parameters	Total GFLOPs
DeepVS	N/A	819.5
STSANet	N/A	493
Casp-Net	51.62 M	283.35
DiffSal	76.57 M	834
DTFSal(V)	40.73 M	259.57
DTFSal	49.08 M	297.91
UniformerSal-S (Ours)	392.7 M	122.2
UniformerSal-ST (Ours)	443.2 M	210.8

B.4.1. Evaluation Metrics

Our comprehensive evaluation employed multiple well-established metrics following recommendations by [3]. Models were trained to minimize negative log-likelihood, with Information Gain prioritized for evaluation as it offers a principled measure of predictive power. **Primary Metric:**

- **Information Gain (IG)** [2]: IG quantifies the informational advantage a model’s prediction provides about human fixation locations over a baseline prior, typically the dataset-specific center bias. It is measured in bits per fixation, with higher values indicating better performance and a more accurate approximation of the true fixation density.

Secondary Metrics:

- **Area Under ROC Curve (AUC-Judd)** [1]: Evaluates the saliency map’s ability to discriminate between fixated and non-fixated locations by treating the map as a binary classifier across varying thresholds. Higher AUC-Judd scores signify better discrimination.
- **Normalized Scanpath Saliency (NSS)** [6]: Calculates the average normalized saliency value at actual human fixation locations. A positive NSS score indicates that fixated locations have higher saliency values than average, with larger scores being better.
- **Pearson’s Correlation Coefficient (CC)**: Measures the linear correlation between the continuous predicted saliency map and the ground truth fixation density map

(typically a blurred map of human fixations). Values closer to 1 indicate stronger positive correlation.

- **Kullback-Leibler Divergence (KLD)**: Measures the divergence between the predicted saliency distribution and the ground truth fixation density distribution. Lower KLD values are preferable, indicating greater similarity between the two distributions.
- **Similarity (SIM)**: Measures the histogram intersection between the predicted saliency map and the ground truth fixation density map. Higher values denote greater similarity.

Unless stated otherwise, scores are computed per frame and then averaged across all test frames within each video, and subsequently averaged across all videos within a given dataset.

B.4.2. Baseline Definitions

Centerbias: For each dataset, a centerbias model was pre-computed by aggregating all human fixation locations from its training set and then applying Gaussian blurring. This model serves as a fundamental lower-bound reference, indicating performance achievable by merely predicting a general tendency to look towards the center. We include the pre-computed centerbias in our model and learn a mixing weight during training. This follows the standard approach in saliency prediction [4].

Gold Standard: This represents an estimate of the upper bound of achievable performance, derived from inter-

observer consistency using a leave-one-subject-out cross-validation approach on the human gaze data, as described by [3] and [8]. The Gold Standard indicates the theoretical maximum performance achievable given the inherent variability in human gaze patterns.

B.4.3. Baseline Model Comparisons

Our comparisons with recent state-of-the-art models were guided by architectural relevance and availability constraints. We focus primarily on TempVST [5] as the most relevant pure transformer-based approach for video saliency prediction. TempVST repurposes the Visual Saliency Transformer (VST) for video fixation prediction by incorporating Divided Space-Time Attention (DSTA), making it the most comparable architecture to our pure transformer approach. While TMAI-Net [7] shows competitive performance, it employs a hybrid architecture combining 3D CNN backbones with Transformer blocks rather than a pure Transformer approach like our UniformerSal-ST. This architectural difference makes it less directly comparable to our work, which specifically investigates pure transformer capabilities for temporal saliency modeling. Both TempVST and TMAI-Net codebases and weights were unavailable at the time of this work. Comparisons against TempVST were limited to publicly reported scores. We primarily refer to their publicly reported scores, acknowledging that direct comparability can be affected by variations in training methodologies, dataset versions, and evaluation protocols. Where TempVST results are available in our comparison tables, they demonstrate the continued challenges in effectively leveraging temporal information, as TempVST’s reported scores on LEDOV did not consistently surpass the static DeepGazeMR baseline across all key metrics.

B.5. Systematic Video Selection Methodology and Analysis

We developed a comprehensive methodology for identifying videos that demonstrate temporal processing effects across different scenarios and dynamics.

B.5.1. Temporal Rescue Analysis

We introduce a novel frame-level analysis method that identifies specific temporal moments where our spatio-temporal model provides computational advantage. This “temporal rescue analysis” employs a dual-marker system:

- **Black dots:** Mark “Metabenchmark” frames where DeepGazeMR fails significantly, defined as having > 1 bit Information Gain (IG) gap from Gold Standard performance.
- **Green dots:** Mark “solved problems” where UniformerSal-ST rescues performance by reducing the gap to < 1 bit IG

This methodology enables temporal mapping of model im-

provements, linking performance gains to specific visual events rather than relying solely on aggregate statistics. **Video Ranking Algorithms.** We employed multiple systematic approaches to rank videos:

1. **Baseline Difficulty Ranking:** Videos ranked by median Information Gain difference between Gold Standard and DeepGazeMR, identifying which videos are most challenging for static models.
2. **Temporal Benefit Ranking:** Videos ranked by conservative temporal benefit calculation. To ensure robust assessment of temporal contributions, we used a conservative approach comparing UniformerSal-ST against the better-performing of either DeepGazeMR or our spatial-only UniformerSal-S variant:

$$\min(\text{GS-DGMR}, \text{GS-S}) - (\text{GS} - \text{ST}) \quad (1)$$

This prevents inflated temporal benefits from architectural improvements unrelated to temporal processing. Larger positive values indicate greater reduction in explanatory gap attributable to temporal processing.

3. **Skewness-based Selection:** Videos ranked by mean-minus-median Information Gain differences to identify cases with particularly variable temporal benefits.

The top 30 videos from each ranking criterion were systematically analyzed, providing comprehensive coverage of different temporal processing scenarios. Figures 11 and 12 demonstrate the systematic identification of videos with the strongest temporal processing benefits, while Figs. 17 and 18 reveal comparative performance patterns across different model variants.

B.6. Frame-Level Window Identification

To systematically identify the ($\text{ST} \gg \text{S}$ and $\text{S} \gg \text{ST}$ windows) performance difference windows, we employed a rigorous quantitative procedure, focusing specifically on scenes from the Metabenchmark. This method involved two distinct analyses: one to find segments where UniformerSal-ST was markedly superior to UniformerSal-S, and a parallel analysis for the converse scenario where UniformerSal-S demonstrated superiority over UniformerSal-ST. The core of this procedure was a frame-level eligibility check using Information Gain (IG) relative to Gold Standard (GS) human gaze data. Specifically, when identifying segments where one model outperformed the other, the better-performing model’s IG had to be higher than the lesser-performing model’s IG by a substantial relative margin. This margin was varied across different runs of the analysis; requiring the better model’s IG to exceed the lesser model’s IG by at least 20%, 50%, 75%, or 90% of the Gold Standard’s IG. Following this frame-level assessment for each video, we identified all uninterrupted sequences of consecutive eligible frames. Only those sequences meeting or exceeding a continuous duration

of at least 5 frames were retained. For each such qualifying window, we recorded its start and end frames/times. We also calculated a summary score for the window, representing the average ratio of the performance difference between the better-performing model and the lesser-performing model, to the performance difference between the Gold Standard and the better-performing model, across the frames in that window. This score effectively quantifies the superior model’s improvement over its counterpart in the context of the remaining gap to Gold Standard performance.

C. Expanded Quantitative Results

C.1. Dataset-Specific Performance Analysis

Our analysis revealed important nuances in how UniformerSal-ST leverages temporal information across different datasets, with clear implications for understanding when temporal processing provides benefits versus challenges.

C.1.1. LEDOV Performance Trends

On LEDOV datasets and their Metabenchmark subsets, UniformerSal-ST demonstrated consistent advantages over UniformerSal-S, with IG improvements of 14.22% (Ledov Test), 15.00% (Ledov Validation), 19.57% (Ledov Test MB), and 25.66% (Ledov Validation MB). These substantial improvements on Metabenchmark subsets, specifically designed to isolate temporal complexities, underscore the significant benefits derived from temporal processing for these video characteristics.

C.1.2. DIEM Performance Challenges

An unexpected but important finding emerged on DIEM datasets, known for dynamic content and frequent camera cuts. Here, UniformerSal-ST underperformed relative to UniformerSal-S, exhibiting IG decreases of -7.52% (DIEM) and -4.79% (DIEM MB). This suggests that certain characteristics of DIEM content—potentially related to rapid scene changes, professional editing, or prevalent camera work—pose challenges for our current temporal modeling approach. These findings highlight that while temporal information is generally beneficial, its precise impact varies significantly depending on video content characteristics, leading to differentiated performance across datasets with different temporal dynamics.

C.2. Detailed Ablation Results

To understand individual component contributions within UniformerSal-ST, we analyzed additional ablation variants beyond the primary UniformerSal-ST versus UniformerSal-S comparison. Evaluations on the combined Metabenchmark revealed a clear performance hierarchy based on Information Gain scores:

- UniformerSal-ST (full model): 1.0528 bits

- UniformerSal $\checkmark\times$ (encoder temporal + no STI): 0.9978 bits
- UniformerSal $\times\checkmark$ (no encoder temporal + STI): 0.9866 bits
- UniformerSal-S (spatial-only): 0.9467 bits

This hierarchical performance (UniformerSal-S < {UniformerSal $\times\checkmark$, UniformerSal $\checkmark\times$ } < UniformerSal-ST) demonstrates additive benefits from both the encoder’s global temporal context and the STI module’s effective integration. Both components contribute significantly to final performance, with the complete model achieving optimal results on temporally demanding scenarios.

D. Extended Instructive Examples

D.1. Reproduction of [8]’s examples with Temporal Rescue Analysis

We reproduced additional key examples from the [8] paper while adding our temporal rescue analysis to provide frame-level insights into when and why temporal processing succeeds or fails (Fig. 1). The systematic analysis of these established challenging examples demonstrates clear progress in temporal saliency modeling while revealing specific moments where improvements occur. The temporal rescue analysis on these videos shows that UniformerSal-ST successfully addresses many of the specific failure modes identified by [8], with green dots indicating frames where temporal processing rescues performance from baseline failures (Figs. 2 and 3).

D.2. Extended Examples

D.2.1. Tangemann Examples

Analysis of additional challenging examples previously highlighted by [8] shows clear improvements:

- **human_gymnastics07**: UniformerSal-ST better captures observers’ typical fixation on the gymnast’s torso while DeepGazeMR often highlights hands or static background elements (Fig. 4)
- **manmade_robot12**: Complex dishwasher loading interactions where some challenging dynamics remain difficult for temporal processing (Fig. 5)
- **manmade_truck05**: Where people appear peripherally during camera pursuit, though improvement over DeepGazeMR is modest, potentially confounded by camera-induced motion (Fig. 6)

D.2.2. Further Temporal Successes

Beyond the examples selected by [8], UniformerSal-ST demonstrates broader capabilities:

- **50_people_london_no_voices**: Better prediction of speaker transitions in interview-style content (Fig. 7)
- **documentary_dolphins**: Effective handling of combined object and camera motion during diving sequences (Fig. 8)

	Model	IG \uparrow	AUC \uparrow	CC \uparrow	KLD \downarrow	NSS \uparrow	SIM \uparrow
L-Val	Centerbias	0.0	0.83	0.34	1.76	1.56	0.28
	GoldStandard	1.92	0.92	0.38	2.36	4.94	0.22
	DeepGazeMR	1.41	0.91	0.67	1.0	3.85	0.52
	TempVST	N/A	N/A	N/A	N/A	N/A	N/A
	UniformerSal $\times\times(S)$	1.45	0.92	0.67	1.0	3.98	0.53
	UniformerSal $\times\checkmark$	1.6	0.92	0.7	0.93	4.24	0.55
	UniformerSal $\checkmark\times$	1.51	0.92	0.68	0.96	4.03	0.53
L-Val MB	UniformerSal $\checkmark\checkmark(ST)$	1.67	0.92	0.71	0.88	4.27	0.55
	Centerbias	0.0	0.81	0.3	2.04	1.44	0.25
	GoldStandard	2.56	0.94	0.41	2.13	5.39	0.23
	DeepGazeMR	1.08	0.9	0.53	1.48	3.12	0.41
	UniformerSal $\times\times(S)$	1.19	0.91	0.55	1.44	3.35	0.43
	UniformerSal $\times\checkmark$	1.38	0.92	0.57	1.35	3.55	0.45
	UniformerSal $\checkmark\times$	1.25	0.91	0.55	1.39	3.38	0.43
L-Test	UniformerSal $\checkmark\checkmark(ST)$	1.5	0.92	0.57	1.27	3.64	0.45
	Centerbias	0.0	0.84	0.35	1.68	1.59	0.29
	GoldStandard	1.8	0.92	0.37	2.37	4.64	0.21
	DeepGazeMR	1.35	0.92	0.68	0.93	3.65	0.52
	TempVST	N/A	0.906	0.725	1.016	3.143	N/A
	UniformerSal $\times\times(S)$	1.43	0.92	0.68	0.92	3.87	0.54
	UniformerSal $\times\checkmark$	1.57	0.93	0.72	0.86	4.14	0.57
L-Test MB	UniformerSal $\checkmark\times$	1.5	0.92	0.71	0.86	4.0	0.55
	UniformerSal $\checkmark\checkmark(ST)$	1.63	0.93	0.73	0.82	4.25	0.57
	Centerbias	0.0	0.83	0.32	1.86	1.51	0.27
	GoldStandard	2.32	0.94	0.39	2.18	4.99	0.22
	DeepGazeMR	0.87	0.89	0.51	1.38	2.74	0.41
	UniformerSal $\times\times(S)$	1.23	0.91	0.58	1.21	3.36	0.46
	UniformerSal $\times\checkmark$	1.41	0.92	0.62	1.12	3.65	0.5
DIEM	UniformerSal $\checkmark\times$	1.33	0.92	0.61	1.14	3.52	0.47
	UniformerSal $\checkmark\checkmark(ST)$	1.47	0.92	0.63	1.08	3.65	0.49
	Centerbias	0.0	0.89	0.44	1.42	2.05	0.35
	GoldStandard	1.53	0.94	0.36	2.36	4.65	0.2
	DeepGazeMR	0.66	0.91	0.61	1.01	3.11	0.48
	UniformerSal $\times\times(S)$	0.9	0.93	0.66	0.88	3.44	0.53
	UniformerSal $\times\checkmark$	0.79	0.93	0.65	0.92	3.31	0.52
DIEM MB	UniformerSal $\checkmark\times$	0.91	0.93	0.66	0.87	3.42	0.52
	UniformerSal $\checkmark\checkmark(ST)$	0.84	0.93	0.65	0.9	3.35	0.51
	Centerbias	0.0	0.87	0.39	1.61	1.85	0.32
	GoldStandard	1.93	0.95	0.37	2.24	4.92	0.21
	DeepGazeMR	0.13	0.89	0.41	1.5	2.06	0.35
	UniformerSal $\times\times(S)$	0.71	0.92	0.54	1.18	2.86	0.44
	UniformerSal $\times\checkmark$	0.63	0.92	0.54	1.2	2.78	0.44
MB	UniformerSal $\checkmark\times$	0.74	0.92	0.55	1.16	2.87	0.43
	UniformerSal $\checkmark\checkmark(ST)$	0.68	0.92	0.54	1.18	2.8	0.43
	Centerbias	0.0	0.85	0.35	1.76	1.68	0.29
	GoldStandard	2.16	0.95	0.38	2.2	5.04	0.22
	DeepGazeMR	0.52	0.89	0.46	1.47	2.45	0.38
	UniformerSal $\times\times(S)$	0.94	0.91	0.55	1.24	3.09	0.44
	UniformerSal $\times\checkmark$	0.98	0.92	0.57	1.22	3.16	0.46
	UniformerSal $\checkmark\times$	0.99	0.92	0.56	1.2	3.14	0.44
	UniformerSal $\checkmark\checkmark(ST)$	1.05	0.92	0.57	1.18	3.19	0.45

Table 3. Performance comparison on key datasets. Our primary comparison is between the spatial-only model (UniformerSal-S, denoted by $\times\times$) and the full spatio-temporal model (UniformerSal-ST, denoted by $\checkmark\checkmark$). We also show ablations without the encoder’s temporal block ($\times\checkmark$) and without the STI module ($\checkmark\times$).

- **game_trailer_ghostbusters:** Sensitivity to dynamic lighting changes and transient salient elements (Fig. 9)
- **news_us_election_debate:** Effective handling of camera zoom-outs while maintaining focus on the subjects. (Fig. 10)

D.2.3. Systematically Selected Examples

Our quantitative ranking procedures identified additional success cases where temporal processing provides the largest measurable benefits (Fig. 11). These systematically selected examples, ranked by conservative temporal benefit calculations, provide unbiased evidence of temporal processing advantages across diverse video content. Videos selected

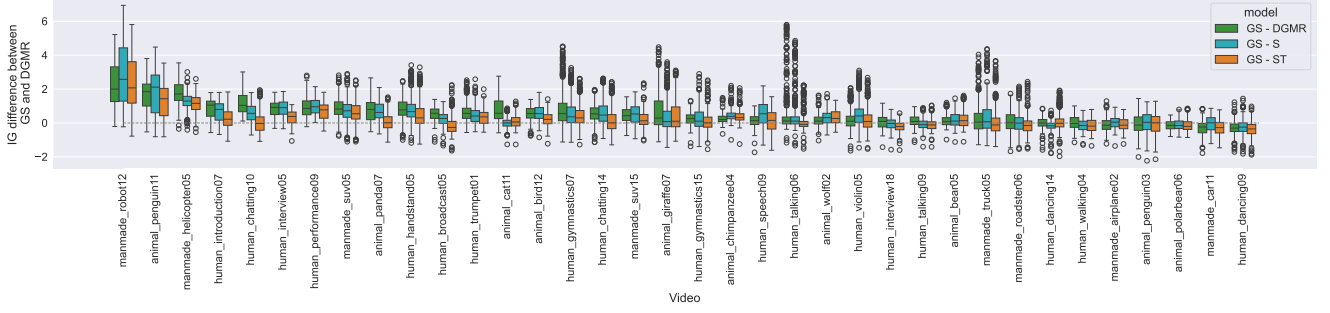


Figure 1. Performance comparison on the video examples selected in [8], showing Information Gain differences between Gold Standard and DeepGazeMR, our spatial-only model (UniformerSal-S), and spatio-temporal model (UniformerSal-ST). This reproduces the challenging examples from [8] while demonstrating our models’ performance across the same test cases.

through skewness-based criteria reveal cases with particularly variable temporal benefits (Fig. 12), identifying scenarios where temporal processing provides intermittent but substantial improvements during specific temporal segments.

D.3. Extended Failure Analysis

Understanding failure modes is crucial for interpreting model behavior and guiding future improvements. In addition to the hand-object interaction bias, we observe other types of failures.

D.3.1. Camera Motion Challenges

Certain camera motions pose specific challenges for temporal processing:

- **50_people_london,** **documentary_coral_reef_adventure:** Rapid camera zoom-ins where UniformerSal-S provides more stable predictions on expanding central objects (Figs. 13 and 14)
- **sport_football_best_goals:** All-or-nothing motion bias attributing saliency broadly to all moving player groups rather than focusing on lead attacker and goalkeeper (Fig. 15)

Mechanistic Interpretation of Zoom Effects: During rapid zoom-ins, the large-scale, uniform expansive optic flow appears to overwhelm UniformerSal-ST’s temporal processing modules. The rapid exit of peripheral information may be misinterpreted as a highly salient event occurring across large portions of the visual field. In contrast, UniformerSal-S, relying solely on spatial features, remains less perturbed by such global, uniform motion, maintaining focus on spatially distinct regions. Conversely, during zoom-outs, UniformerSal-ST often performs better, leveraging motion continuity to track shrinking central objects even as spatial detail diminishes.

D.3.2. Incorrect Object Focus

Specific instances where temporal processing leads to incorrect generalizations:

- **BBC_wildlife_eagle:** Predominant focus on the deer’s head while spatial models correctly predict entire animal body coverage (Fig. 16)

D.3.3. Systematically Identified Challenges

Our comprehensive ranking analysis identified videos where UniformerSal-S consistently outperforms UniformerSal-ST, providing systematic rather than anecdotal evidence of temporal processing limitations (Fig. 18). These systematically identified failure cases offer crucial insights for understanding when temporal information becomes counterproductive and guide future architectural improvements. The observed patterns of temporal processing challenges provide data points for understanding the conditions under which our approach struggles, informing both current model interpretation and future development directions. Figure 17 reveals cases where even our improved spatial architecture faces challenges, highlighting the strongest opportunities for temporal processing improvements.

E. Temporal Fusion Analysis: Feature Extraction and Statistical Validation

To provide a mechanistic understanding of why temporal fusion fails on DIEM but succeeds on LEDOV, we conducted a comprehensive frame-level analysis using automatically extracted video features. This section details our methodology and key findings.

E.1. Feature Extraction Methodology

We extracted four categories of frame-level features from all videos in DIEM (84 videos, ~214,000 frames) and LEDOV (77 videos across test and validation sets, ~26,200 frames):

1. **Scene Cuts:** Detected using PySceneDetect’s ContentDetector with default threshold (27.0). Cut frames mark the



Figure 2. Temporal dynamics for six key examples selected in [8], using Normalized Scanpath Saliency (NSS). Black dots indicate problem frames where DeepGazeMR fails; green dots show frames where UniformerSal-ST rescues performance. This temporal rescue analysis reveals specific moments where temporal processing provides computational advantage.



Figure 3. Temporal dynamics for six key examples selected in [8], using Information Gain (IG). The temporal rescue analysis reveals specific moments where temporal processing provides computational advantage, with black dots marking baseline failures and green dots indicating successful temporal rescues.

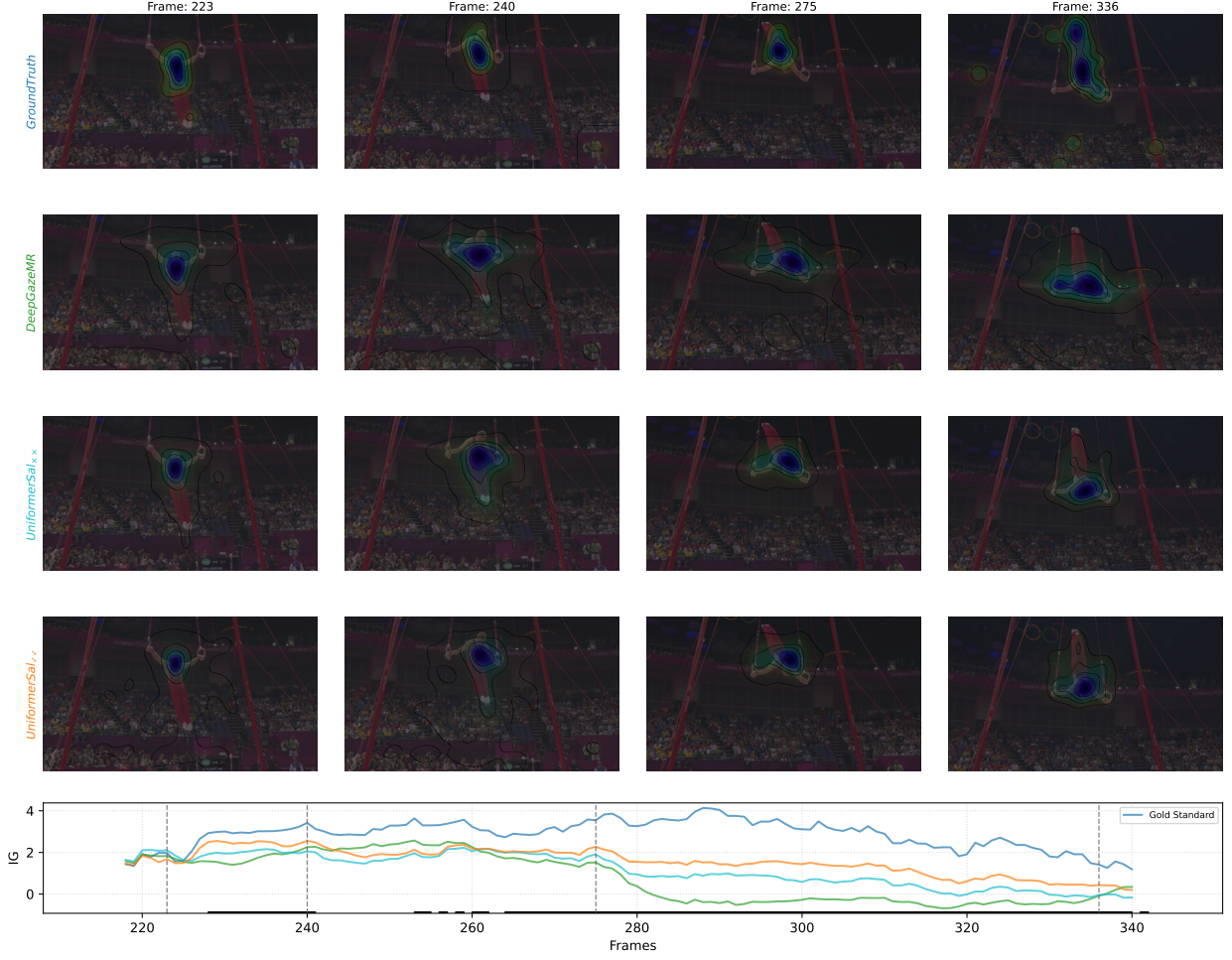


Figure 4. Temporal dynamics in human_gymnastics07 showing UniformerSal-ST’s improved focus on the gymnast’s body compared to DeepGazeMR’s attention to hands and background elements. The x-axis markers indicate that the frame is a Metabenchmark frame.

start of new scenes.

2. **Camera Motion:** Estimated via affine transformation between consecutive frames using sparse optical flow (Lucas-Kanade). We decompose the affine matrix into zoom (scale factor > 1 or < 1), pan (translation magnitude), and rotation components.
3. **Object Motion:** Computed as the residual optical flow after subtracting the expected camera-induced flow field. We report mean, 90th percentile, and spatial coverage of residual flow magnitude.
4. **Semantic Continuity:** Measured using CLIP (ViT-B/32) embeddings extracted at every frame. Frames are labeled “discontinuous” if the minimum cosine similarity between the current frame’s embedding and any frame in its 16-

frame window falls below 0.85.

Since our model processes 16-frame sliding windows, a frame’s prediction is “contaminated” if *any* frame in its preceding window contains the effect of interest (cut, high motion, or semantic discontinuity).

E.2. Dataset Characteristics

Table 4 reveals fundamental differences between DIEM and LEDOV that explain the divergent performance of temporal fusion.

Critically, DIEM has **92× more scene cuts** than LEDOV (9.2% vs. 0.1%) and **38× more semantic discontinuity** (15.2% vs. 0.4%). Conversely, LEDOV has *more* motion than DIEM, yet temporal fusion succeeds there—indicating

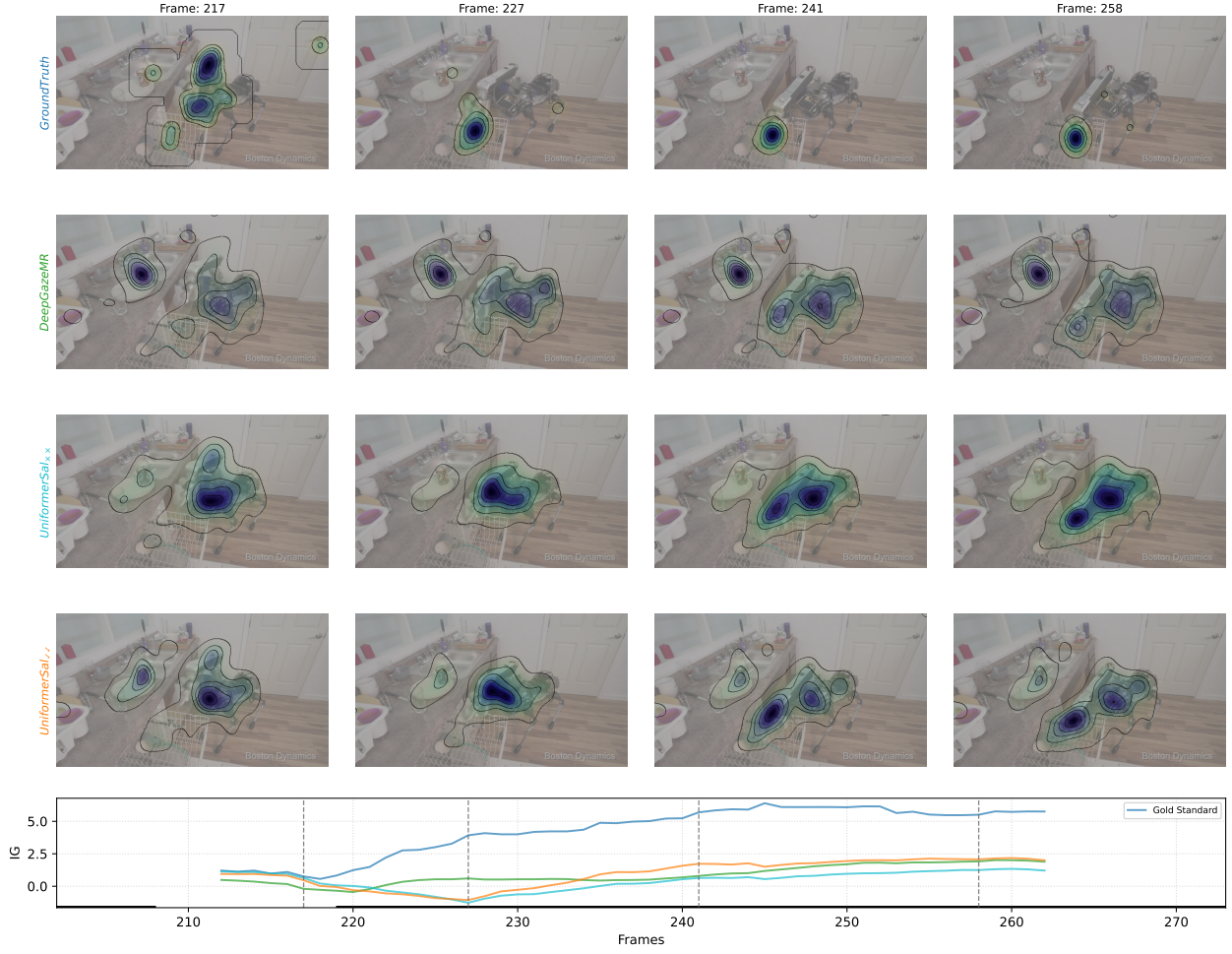


Figure 5. Complex interaction sequence in manmade_robot12 (dishwasher loading) demonstrating challenges that remain difficult even for temporal processing, illustrating limitations of current approaches. The x-axis markers indicate that the frame is a Metabenchmark frame.

Characteristic	DIEM	LEDOV-Test	LEDOV-Val
Scene Cuts	9.2%	0.1%	0.1%
Semantic Discontinuity	15.2%	0.6%	0.4%
Any Camera Motion	52.9%	73.4%	73.4%
Any Object Motion	40.0%	~53%	~54%
Triply Clean Frames	43.7%	22.3%	22.6%

Table 4. Comparison of dataset characteristics. DIEM contains significantly more scene cuts and semantic discontinuities—the specific vulnerabilities of temporal fusion.

Model	Clean IG	Contam. IG	Δ	p -value
UniformerSal-ST	0.849	0.815	+0.034	<0.001
UniformerSal-S	0.949	0.955	−0.006	0.342

Table 5. Impact of scene cuts on model performance (DIEM). Scene cuts significantly degrade UniformerSal-ST while leaving UniformerSal-S unaffected.

E.3. Statistical Analysis: Scene Cuts

We performed Welch’s t-tests on pooled frame-level Information Gain (IG) scores, comparing “clean” frames (no cut in the 16-frame window) to “contaminated” frames (cut present in window).

that motion per se is not the problem.

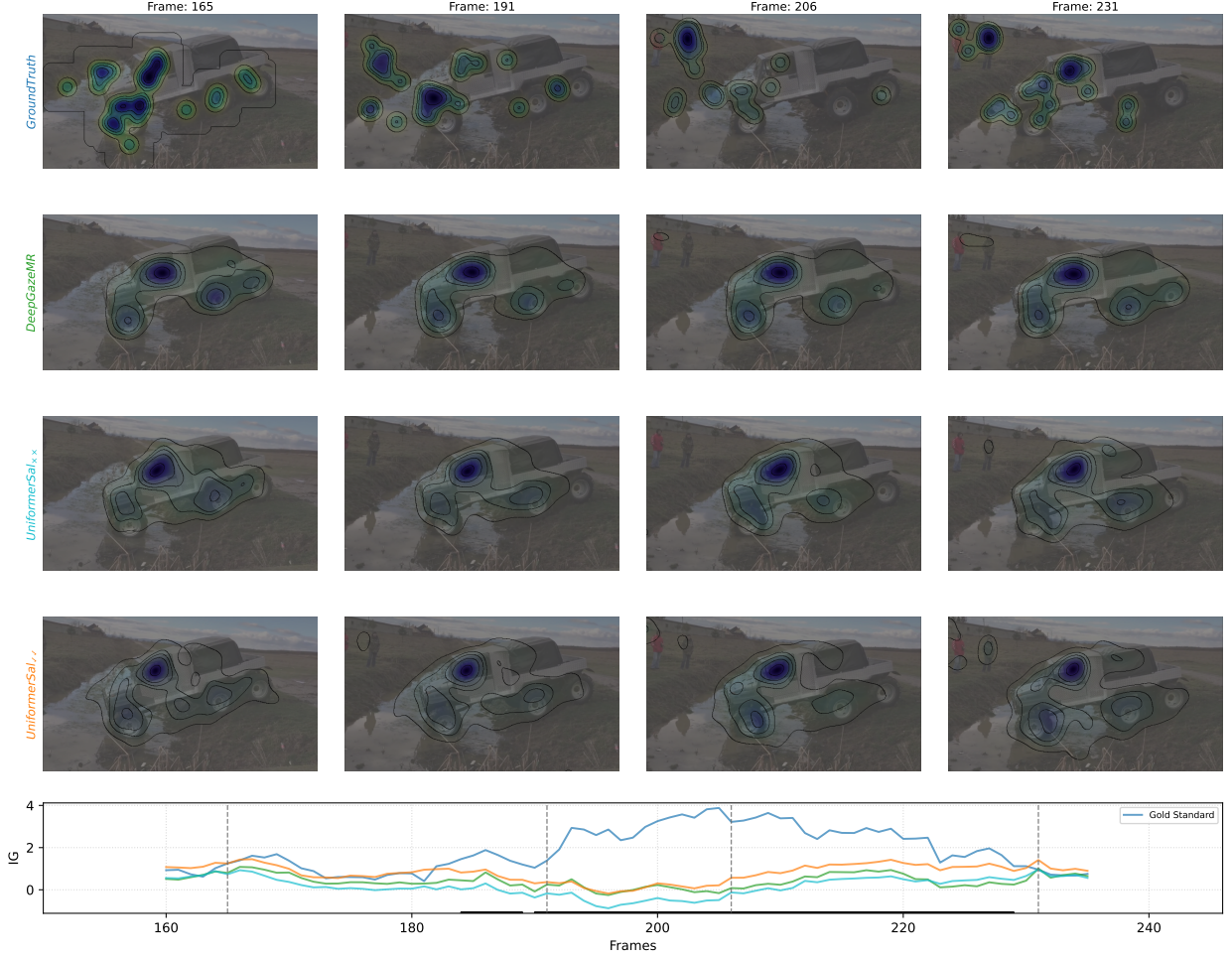


Figure 6. Camera pursuit sequence in manmade_truck05 where peripheral people appear, showing modest temporal improvement potentially confounded by camera-induced motion effects. The x-axis markers indicate that the frame is a Metabenchmark frame.

Table 5 confirms that scene cuts are an **ST-specific vulnerability**: cuts hurt ST by 0.034 bits ($p < 0.001$), while S is statistically unaffected ($p = 0.342$). This validates our mechanistic explanation that cuts “pollute” the global temporal token.

E.4. Combined Analysis: Triply-Clean Frames

To test whether temporal fusion can succeed when *all* confounds are removed, we analyzed “triply-clean” frames: those with no scene cuts, no high motion (below 90th percentile), and semantic continuity (CLIP similarity ≥ 0.85) throughout their 16-frame window.

Table 6 reveals a striking finding: even on DIEM’s triply-clean frames, **S still outperforms ST by 0.105 bits**. This

Dataset	Triply-Clean %	ST (IG)	S (IG)	Winner
DIEM	43.7%	0.838	0.943	S (−0.105)
LEDOV-Test	22.3%	1.686	1.502	ST (+0.184)
LEDOV-Val	22.6%	1.927	1.543	ST (+0.384)

Table 6. Model performance on triply-clean frames. Even under ideal conditions, S outperforms ST on DIEM, while ST excels on LEDOV.

suggests that DIEM’s content dynamics—beyond just editing artifacts—are fundamentally incompatible with our temporal fusion approach. In contrast, ST achieves substantial gains on LEDOV’s triply-clean frames (+0.184 to +0.384 bits), confirming that temporal fusion provides genuine value

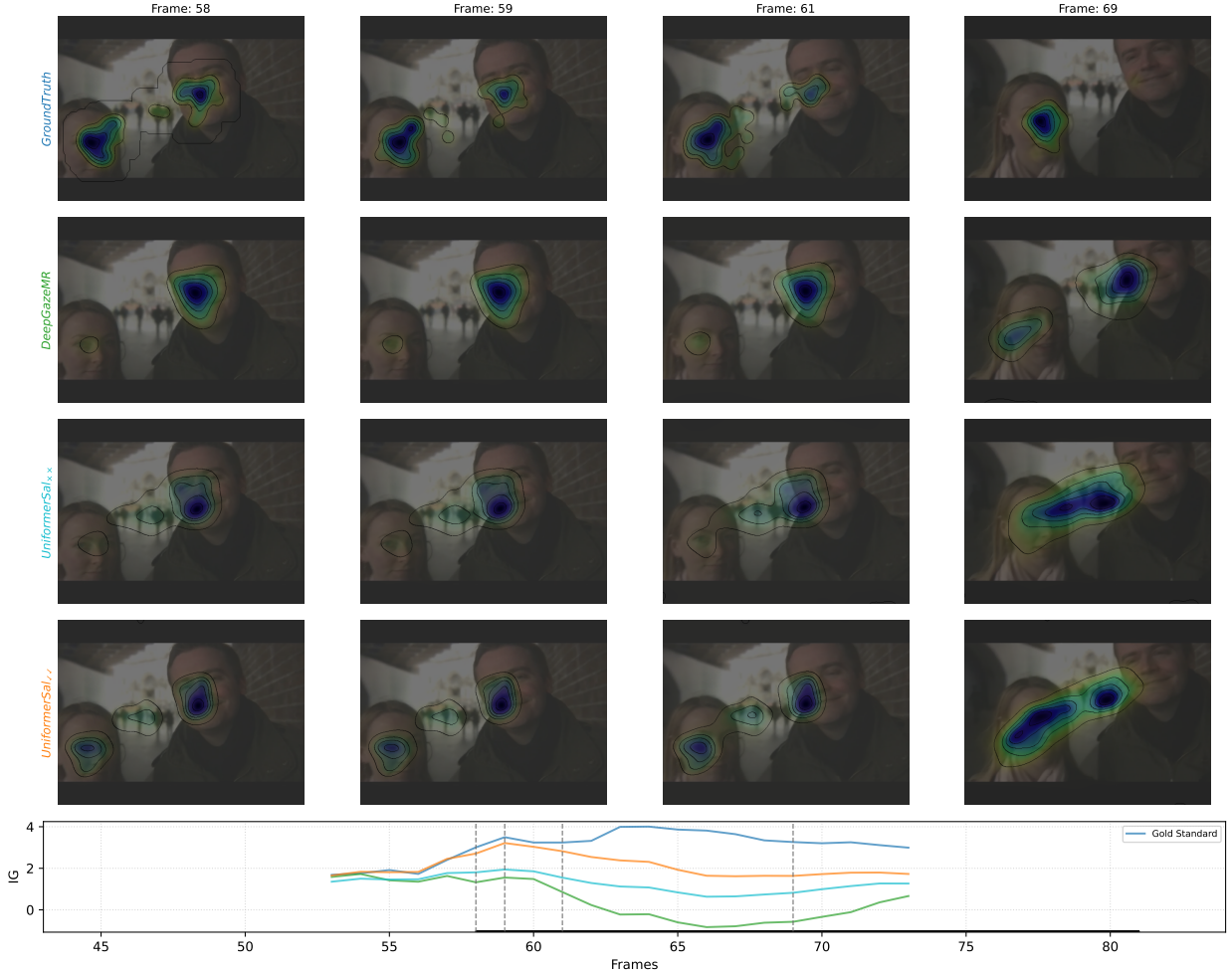


Figure 7. Interview-style content in 50_people_london_no_voices showing UniformerSal-ST’s superior prediction of speaker transitions and maintenance of focus during conversational dynamics. The x-axis markers indicate that the frame is a Metabenchmark frame.

when content characteristics align with the model’s assumptions.

E.5. Limitations

We note that this analysis represents a practical approximation rather than a fully controlled experiment. Ideally, one would construct synthetic datasets where specific effects (scene cuts, camera motion, semantic discontinuity) are systematically introduced or removed while holding all other factors constant. Such controlled stimuli would yield more pronounced effect sizes and eliminate potential confounds from correlated video characteristics. Our approach of partitioning naturally-occurring video frames by detected features provides useful insights but cannot fully disentangle effects that may co-occur in real content. Nevertheless, the

consistent patterns we observe—particularly the ST-specific vulnerability to cuts and semantic discontinuity—provide strong evidence for the mechanistic explanations proposed in the main text.

E.6. Summary of Findings

Our analysis identifies a clear hierarchy of factors that hurt temporal fusion on DIEM:

1. **Semantic discontinuity** is the largest factor ($\Delta = +0.123$ bits, ST 37% more vulnerable than S)
2. **Object motion** hurts both models equally (~ 0.09 bits each)
3. **Scene cuts** are ST-specific ($\Delta = +0.034$ bits for ST, S unaffected)

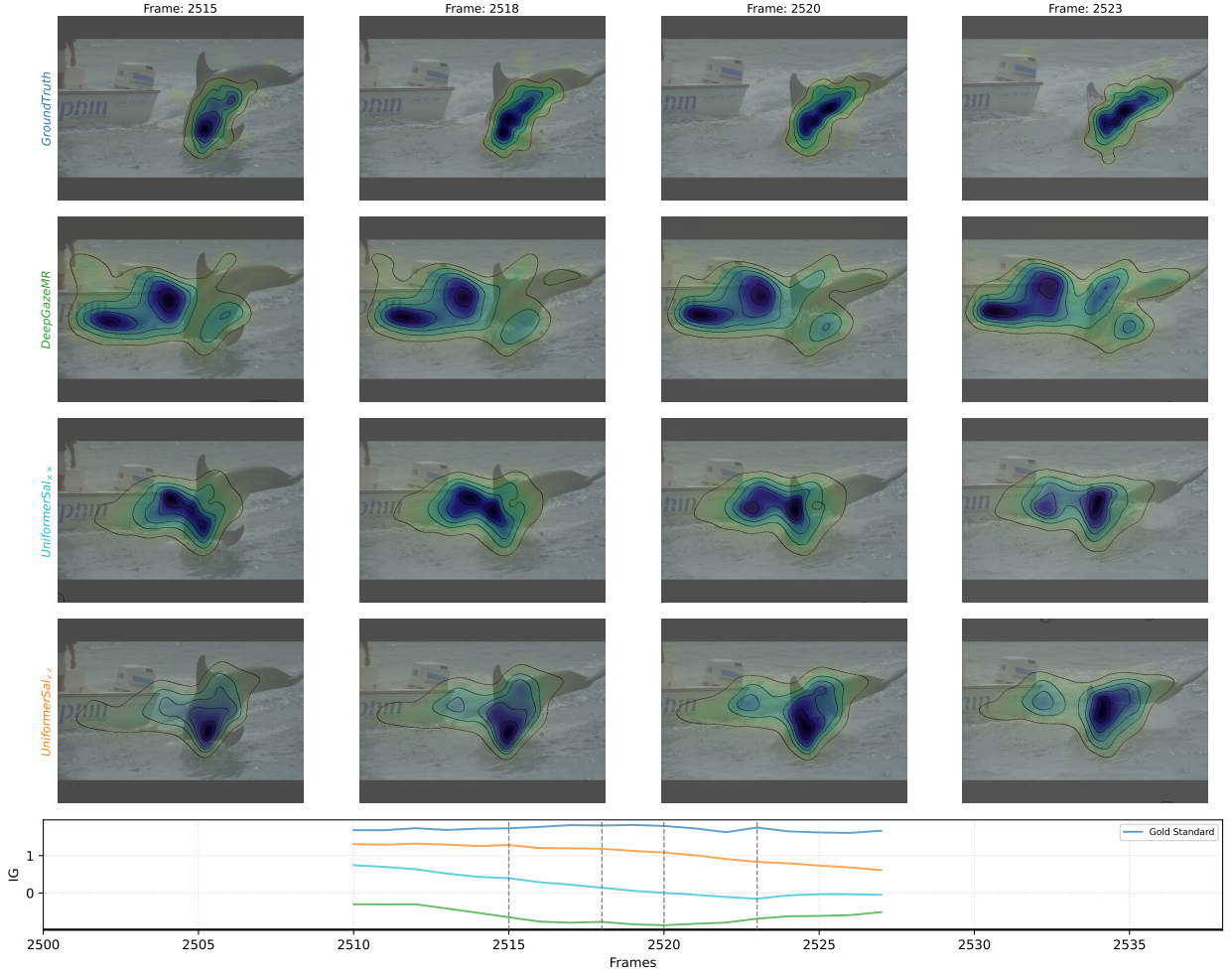


Figure 8. Combined object and camera motion in documentary_dolphins demonstrating effective temporal tracking of dolphin’s diving and complex motion patterns. The x-axis markers indicate that the frame is a Metabenchmark frame.

These findings provide actionable guidance: future temporal fusion methods must incorporate cut detection, adaptive windowing, or mechanisms to distinguish camera-induced from object-induced motion to achieve robust performance across diverse video content.

F. Extended Discussion

Our work set out to resolve the ambiguity surrounding the role of temporal information in video saliency prediction. Our findings suggest a nuanced answer to the question posed in the introduction: on average, temporal information appears less critical than strong static cues, explaining the remarkable performance of static models like DeepGazeMR. However, temporal cues become crucial in specific, dynami-

cally rich scenarios—precisely those where previous video models have failed.

Our findings show that temporal cues, when integrated via a principled fusion mechanism, provide substantial benefits in temporally coherent scenes. However, we also identified a critical failure mode: this same mechanism is vulnerable to scene discontinuities and rapid camera motion prevalent in edited content, leading to performance degradation.

Our results on the Metabenchmark (MB) are particularly telling. The MB was designed to isolate temporally demanding events where static models fail, such as sudden object onsets or complex interactions. The fact that UniformerSal-ST achieves its largest performance gains over

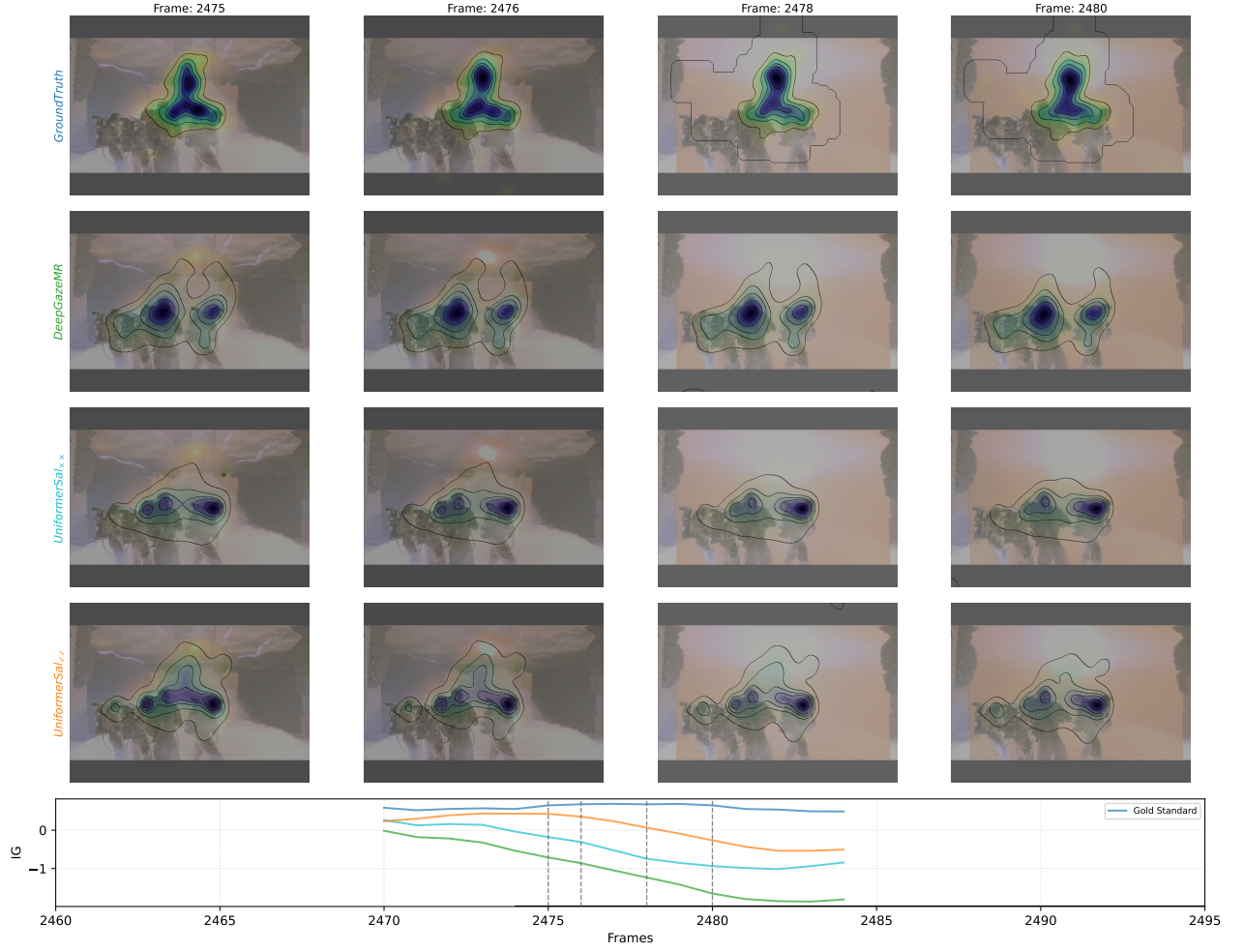


Figure 9. Dynamic lighting changes and transient salient elements in game_trailer_ghostbusters showing temporal sensitivity to sudden visual events and lighting transitions. The x-axis markers indicate that the frame is a Metabenchmark frame.

both DeepGazeMR and UniformerSal-S on the MB subsets provides strong evidence that our temporal fusion mechanism is correctly capturing these critical dynamic events.

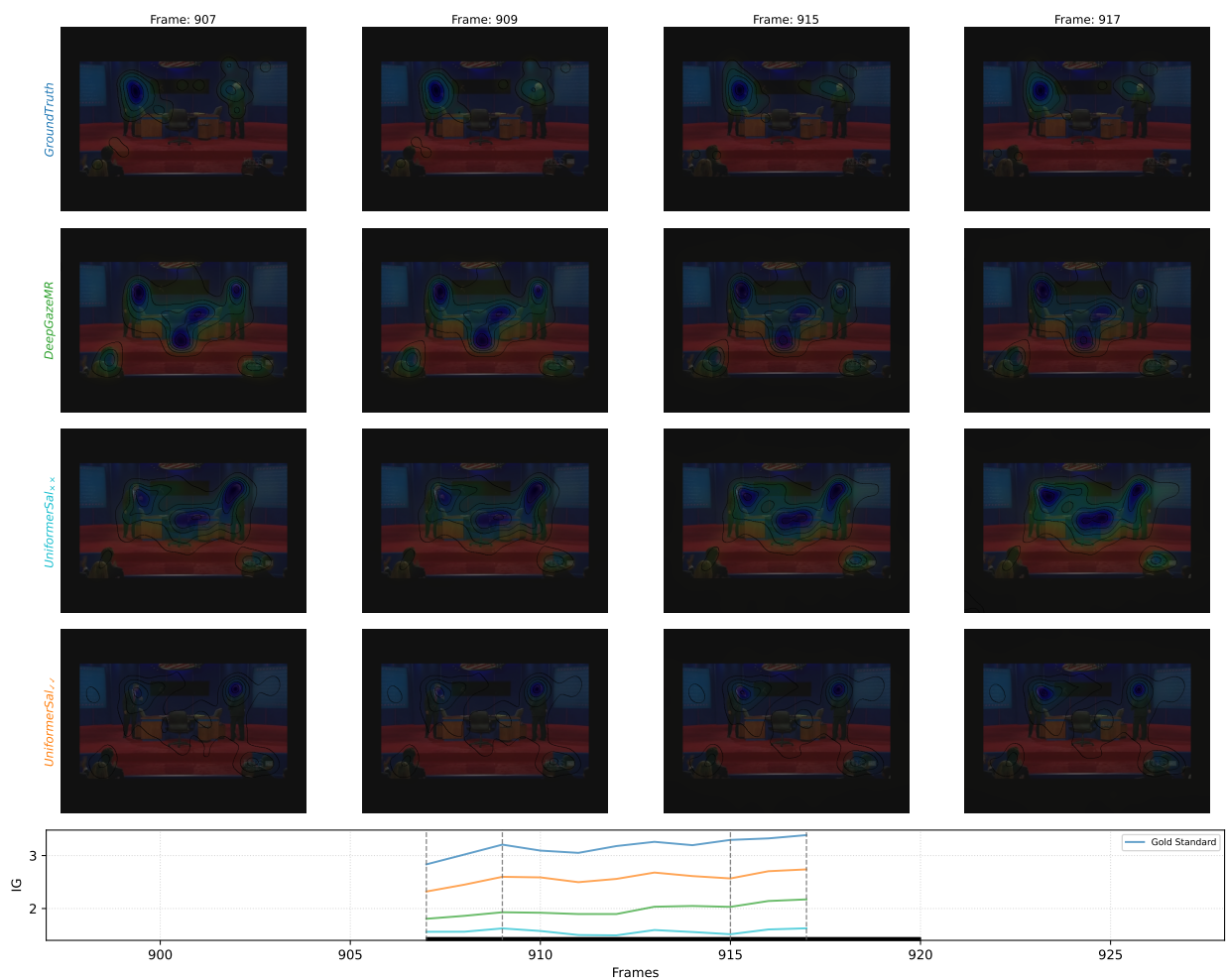


Figure 10. Camera zoom-out sequence in news_us_election_debate demonstrating UniformerSal-ST’s better handling of slight camera zoom-out operations while maintaining focus on the primary subjects. The temporal processing successfully tracks the two people on stage through the zoom transition. The x-axis markers indicate that the frame is a Metabenchmark frame.

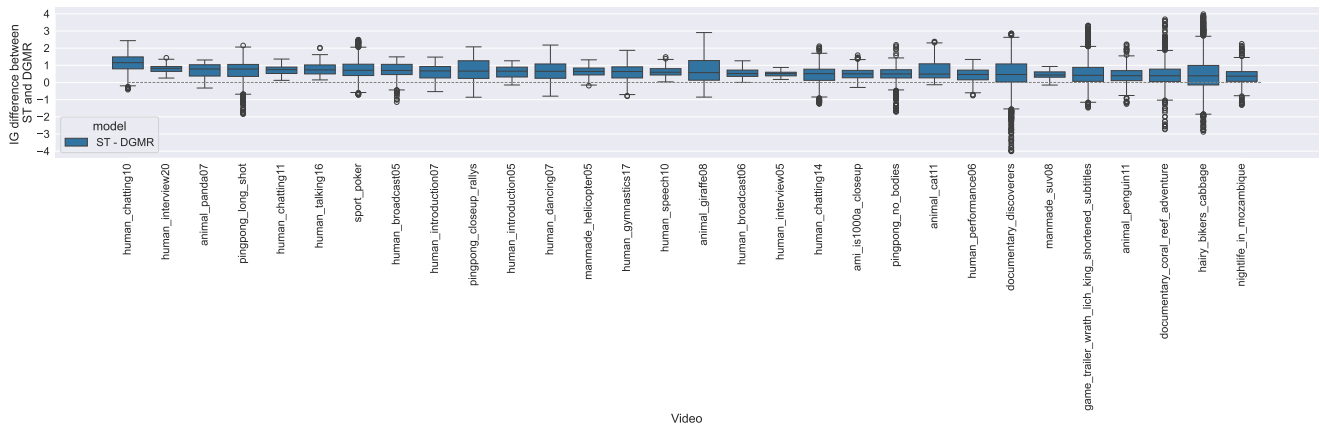


Figure 11. Top 30 videos showing strongest median IG improvement of UniformerSal-ST over DeepGazeMR. These videos demonstrate the largest consistent temporal processing benefits across frames.

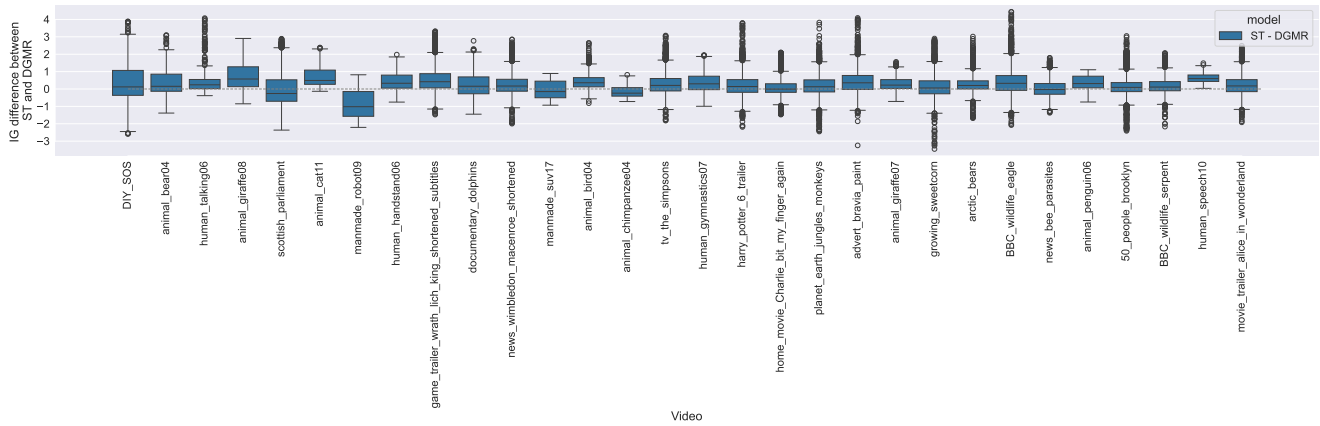


Figure 12. Top 30 videos selected by skewness-based criteria, identifying cases with particularly variable temporal benefits. The mean-minus-median ranking reveals videos where temporal processing provides intermittent but substantial improvements.

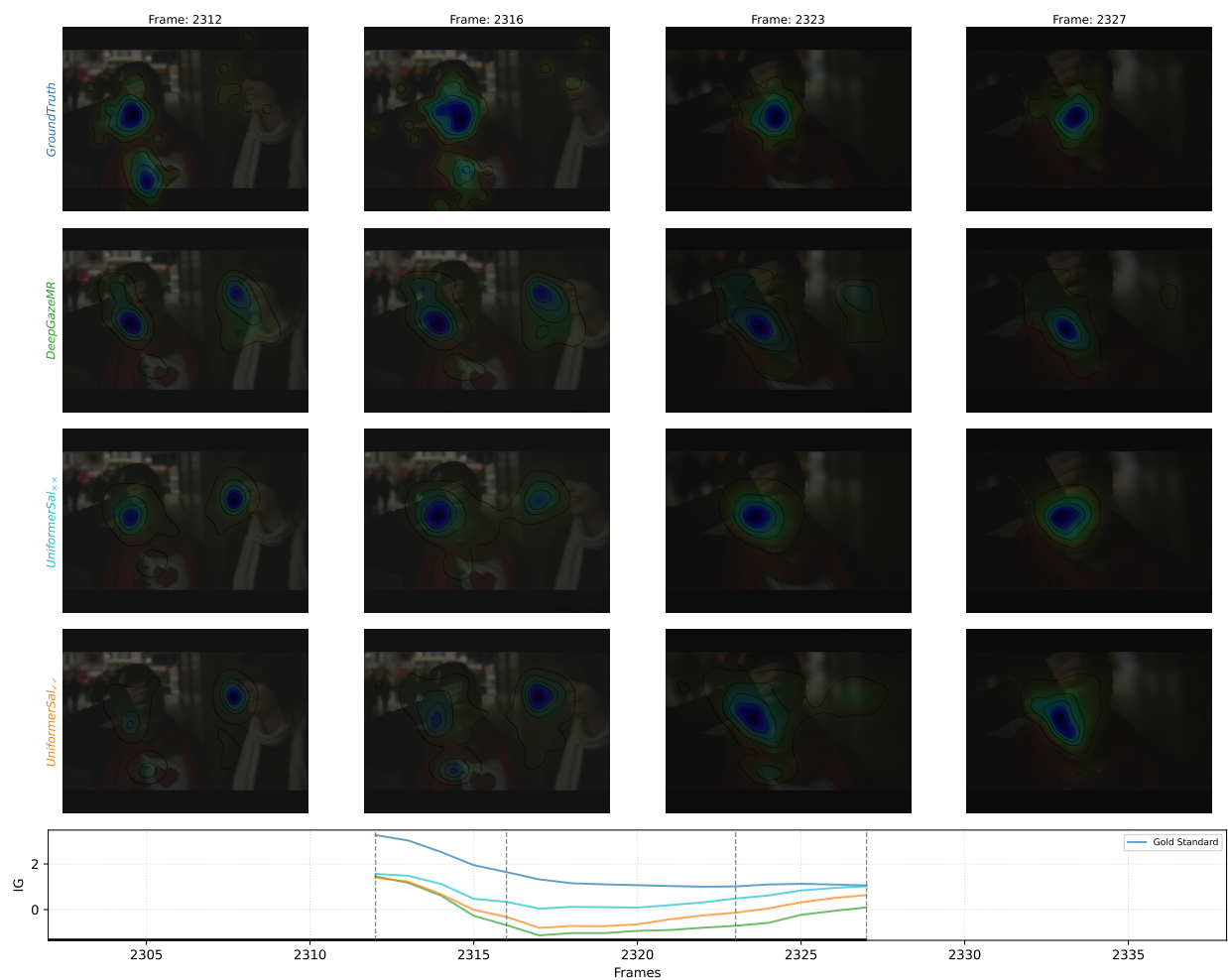


Figure 13. Rapid camera zoom-in sequence in 50_people_london demonstrating temporal processing challenges with global motion. UniformerSal-S provides more stable predictions on the expanding central object, while UniformerSal-ST’s temporal processing appears disrupted by the uniform optical flow during zoom operations. The x-axis markers indicate that the frame is a Metabenchmark frame.

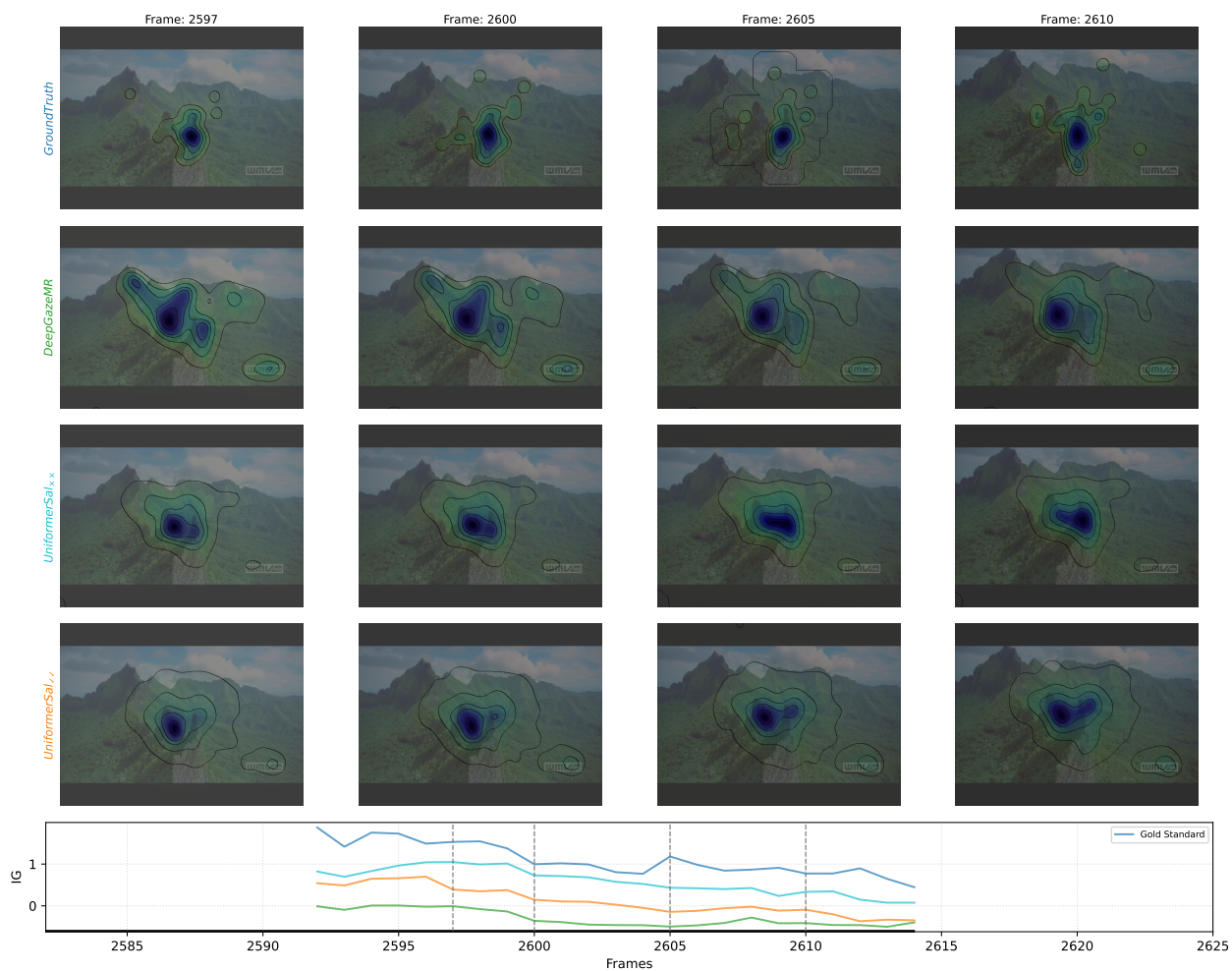


Figure 14. Rapid camera zoom-in sequence in documentary_coral_reef_adventure where UniformerSal-S provides more stable predictions on expanding central objects, demonstrating temporal processing challenges with global motion. The x-axis markers indicate that the frame is a Metabenchmark frame.

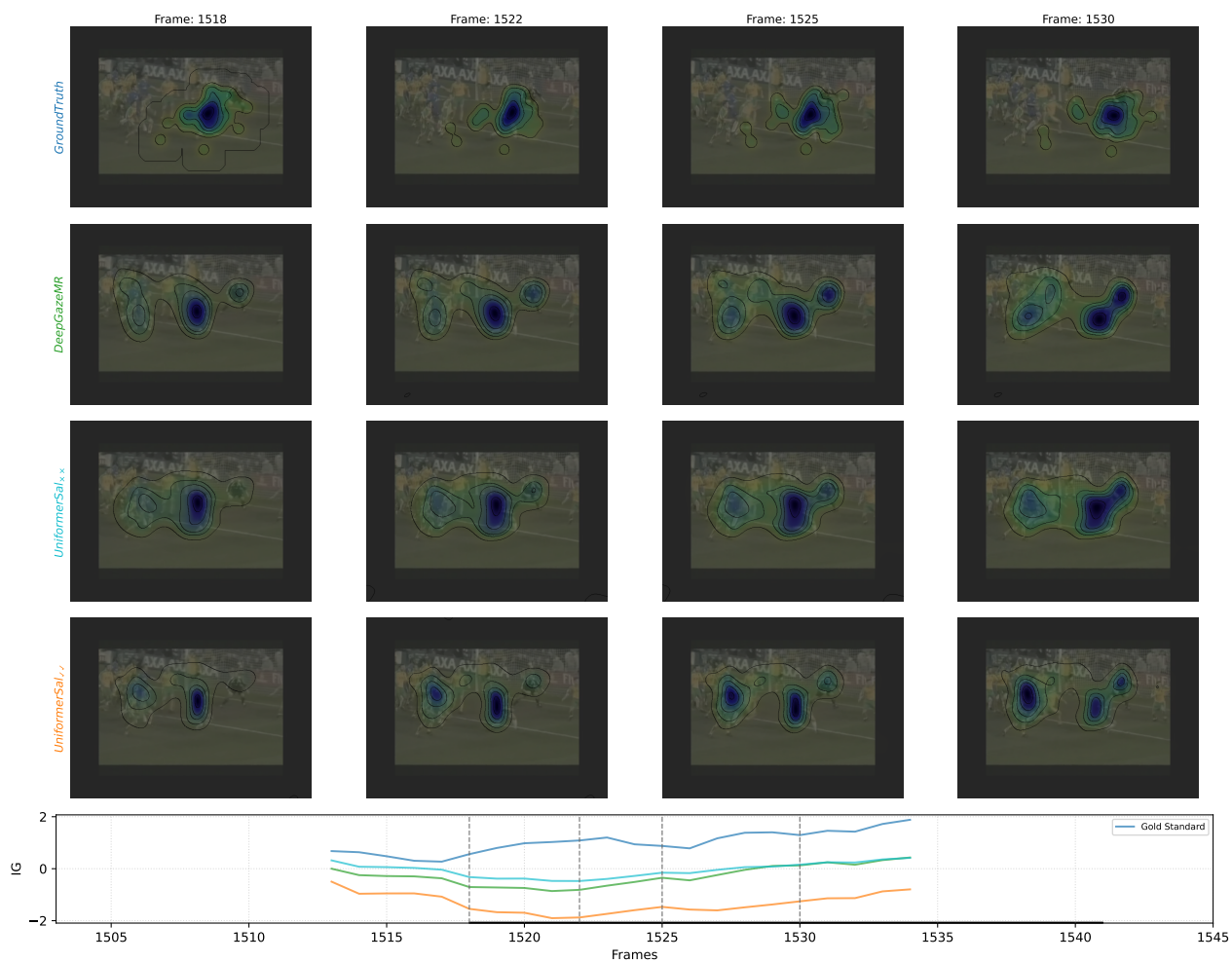


Figure 15. All-or-nothing motion bias in sport_football_best_goals where UniformerSal-ST attributes saliency broadly to all moving player groups rather than focusing on key players, illustrating failure modes with complex multi-object motion. The x-axis markers indicate that the frame is a Metabenchmark frame.

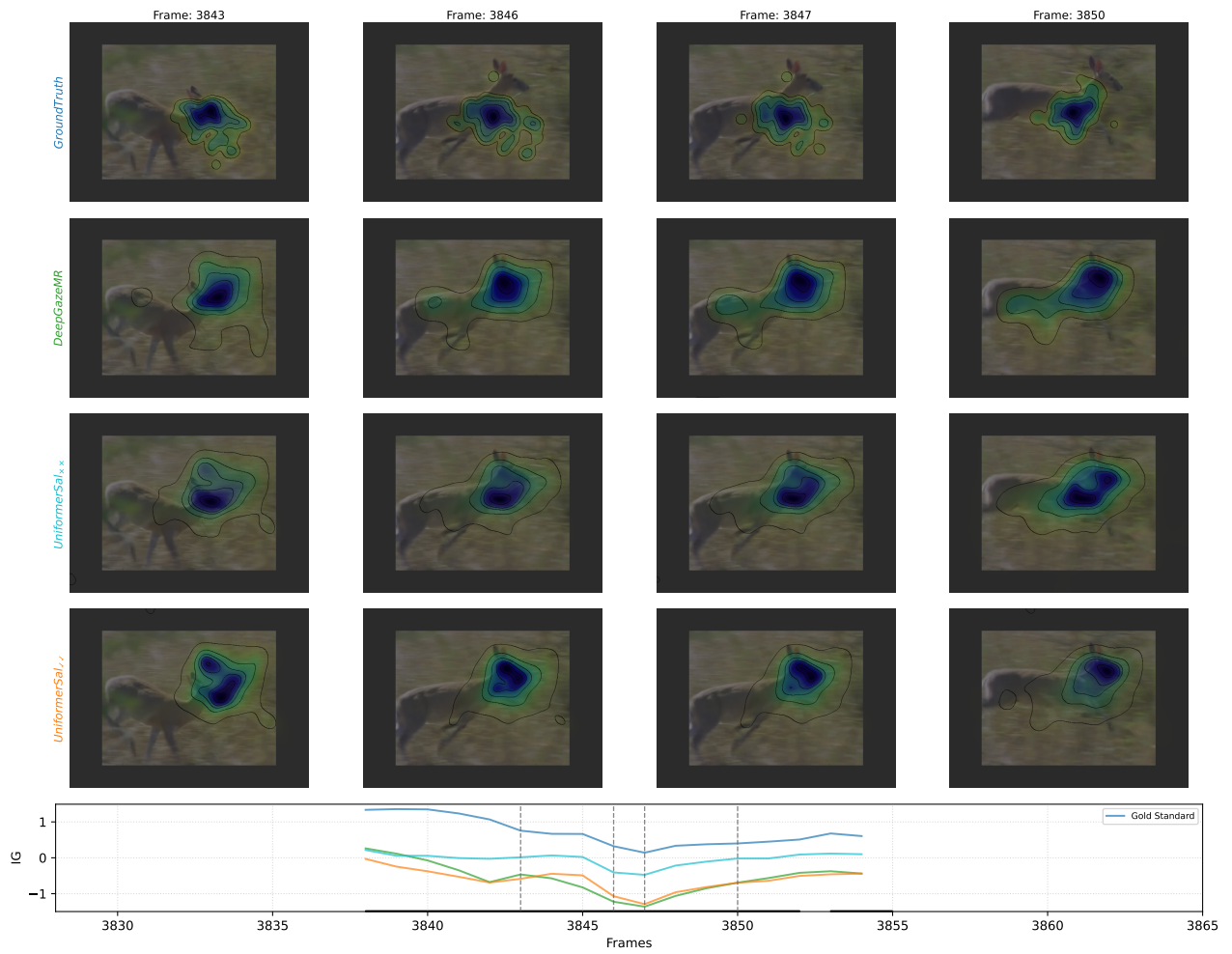


Figure 16. Incorrect focus generalization in BBC_wildlife_eagle where UniformerSal-ST focuses predominantly on the animal’s head while spatial models correctly predict entire body coverage, demonstrating over-specific temporal attention. The x-axis markers indicate that the frame is a Metabenchmark frame.

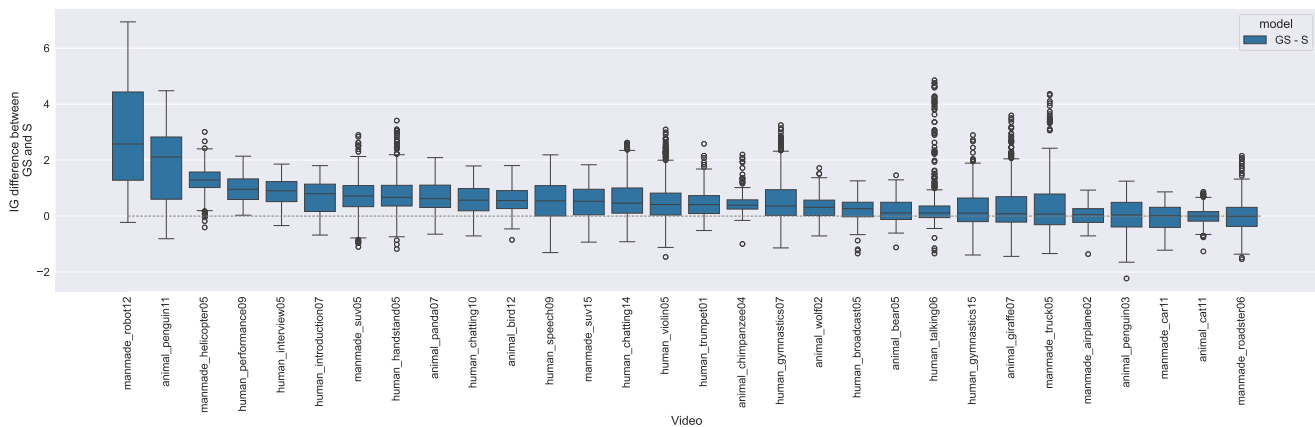


Figure 17. Videos where Gold Standard most exceeds our spatial-only model, identifying remaining temporal challenges even with improved spatial architecture. These cases represent the strongest opportunities for temporal processing improvements.

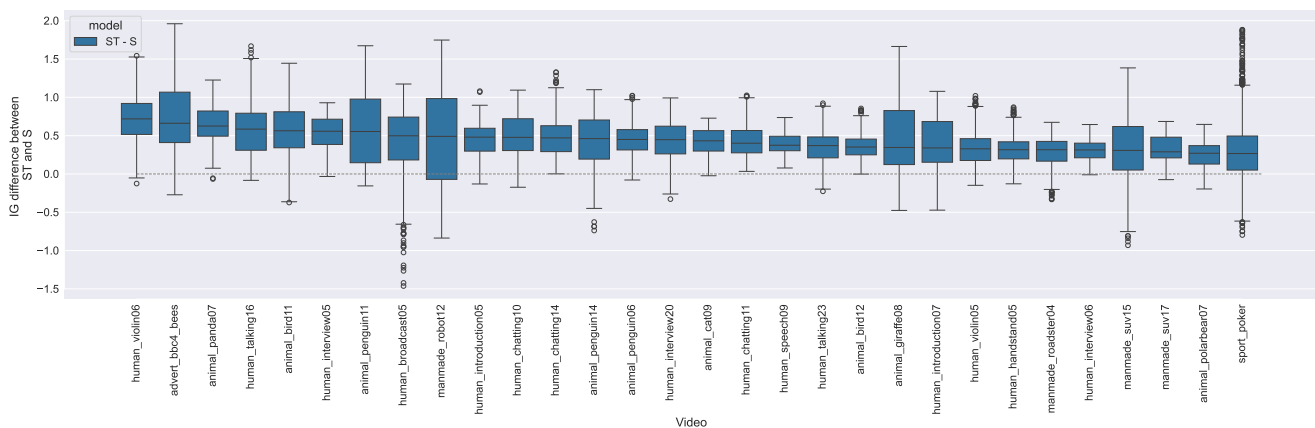


Figure 18. Direct comparison between UniformerSal-ST and UniformerSal-S, isolating pure temporal processing benefits while controlling for spatial architecture. This analysis reveals videos where temporal information provides the most substantial improvements.

References

- [1] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009. 2
- [2] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015. 2
- [3] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3
- [4] Matthias Kümmerer and Matthias Bethge. Predicting Visual Fixations. *Annual Review of Vision Science*, 9(1):269–291, 2023. _eprint: <https://doi.org/10.1146/annurev-vision-120822-072528>. 2
- [5] Nikos Lazaridis, Kostas Georgiadis, Fotis Kalaganis, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Spiros Nikolopoulos, and Ioannis Kompatsiaris. The visual saliency transformer goes temporal: Tempvst for video saliency prediction. *IEEE Access*, 12:129705–129716, 2024. 3
- [6] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005. 2
- [7] Rui Tan, Minghui Sun, and Yanhua Liang. Transformer-based multi-level attention integration network for video saliency prediction. *Multimedia Tools and Applications*, 84(13):11833–11854, 2024. 3
- [8] Matthias Tangemann, Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Measuring the importance of temporal features in video saliency. In *Computer Vision – ECCV 2020: Lecture Notes in Computer Science*, pages 667–684. Springer International Publishing, Cham, 2020. 3, 4, 6, 7, 8
- [9] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. 2021. 1