From Detector Evidence to Language: Explainable Deepfake Video Detection

Supplementary Material

Ablation Study of TRAP and CAVE

In the main paper, TRAP is introduced as an intra-snippet module for modeling localized short-range temporal artifacts, whereas CAVE is designed to model agreement and disagreement across multiple snippets sampled from the same video. To directly validate these intended roles, we perform a dedicated ablation study at $T = 5$ using four detector variants: *Backbone only*, *Backbone+TRAP*, *Backbone+CAVE*, and the full *Backbone+TRAP+CAVE* model.

All ablation variants are trained and evaluated under the same setting as the main detector experiments on FaceForensics++ (c23). Following the main paper, the task is 4-class classification over *Original*, *Face2Face*, *FaceShifter*, and *FaceSwap*, and evaluation is performed on the balanced held-out test split with $T = 5$. This controlled setup ensures that performance differences can be attributed to the inclusion or exclusion of TRAP and CAVE, rather than to changes in data, optimization, or evaluation protocol.

Table 7 shows three clear trends. First, the backbone alone is already a strong baseline and attains the highest Macro-AUC, indicating strong ranking ability even without the proposed temporal modules. Second, TRAP is the stronger standalone addition: compared with the backbone, it preserves overall Acc. while keeping Macro-F1 nearly unchanged, and improves class-wise behavior on manipulations that benefit from stronger local temporal modeling. Third, CAVE alone is weaker in isolation, suggesting that cross-snippet disagreement cues are informative but not sufficiently reliable unless the underlying snippet-level representation is already strong. The full model achieves the best Acc., Macro-F1, and Weighted-F1, showing that TRAP and CAVE are complementary when used jointly.

Table 8 clarifies the role of each module. The backbone alone achieves the highest *Original* accuracy, but not the best *Original* F1. This indicates that it often predicts the *Original* class for manipulated samples, which keeps recall high on real videos but lowers precision. TRAP is the stronger standalone temporal module and is especially helpful on *FaceShifter*, where it substantially improves both Acc. and F1. This is consistent with its intended role of modeling localized short-range temporal inconsistencies within snippets. CAVE alone performs best on *FaceSwap* accuracy, but degrades *Original*, *Face2Face*, and *FaceShifter*, which suggests that cross-snippet disagreement cues can help some manipulations but are not sufficiently stable when used without stronger local temporal features. The full model gives the best overall balance across classes: it achieves the highest F1 on *Original*, *Face2Face*, and *FaceShifter*, while remaining competitive on *FaceSwap*. This pattern supports the intended design of FLARE: TRAP strengthens local snippet-level temporal evidence, and CAVE becomes most effective when applied on top of those stronger representations.

The confusion matrices in Table 9 explain why the full model is better. The backbone-only model strongly overpredicts the *Original* class: 21 *Face2Face*, 14 *FaceShifter*, and 16 *FaceSwap* samples are all misclassified as *Original*. This raises *Original* accuracy but lowers its F1 because precision is reduced. TRAP improves the local temporal representation and clearly reduces some of these fake-to-real confusions, especially for *FaceShifter*, but the detector still keeps a noticeable bias toward predicting *Original*. CAVE alone behaves differently: it is helpful for *FaceSwap*, but it also introduces many real-to-fake errors, including 12 *Original*→*FaceSwap* mistakes, which substantially hurts *Original* performance. The full model gives the best overall trade-off. Compared with the backbone, it reduces *Face2Face*→*Original* errors from 21 to 12 and *FaceShifter*→*Original* errors from 14 to 10, while also increasing the diagonal counts for *Face2Face* from 117 to 126 and for *FaceShifter* from 122 to 129. At the same time, it preserves the stronger local temporal behavior learned by TRAP. In other words, TRAP produces stronger snippet-level temporal features, and CAVE becomes useful only after those features are strong enough for reliable cross-snippet disagreement modeling. This is why the joint model outperforms either standalone module.

Overall, the ablation study supports the intended interpretation of the two modules. TRAP is the more reliable standalone temporal component and is particularly helpful for manipulations with strong local temporal inconsistencies. CAVE alone is weaker and less stable, but becomes beneficial when combined with TRAP. The full FLARE model therefore provides the best balanced classification performance at $T = 5$, even though it does not achieve the best value on every individual class or metric.

Limitations of Detector-Guided Grounding

An important limitation of the proposed detector-guided explanation pipeline is the possibility of error propagation from the detector to the explanation stage. Since the ROI is derived from a detector-side saliency signal, an inaccurate or unstable heatmap may cause the explanation model to attend to the wrong spatial region, even when the generated text itself appears plausible. More generally, if the detector relies on biased or spurious cues, the resulting explanation may inherit these biases, because the language

Model Variant	Acc. ↑	Macro-F1 ↑	Weighted-F1 ↑	Macro-AUC ↑
Backbone only	0.8768	0.8804	0.8804	**0.9832**
Backbone + TRAP	0.8768	0.8795	0.8795	0.9779
Backbone + CAVE	0.8464	0.8480	0.8480	0.9716
Backbone + TRAP + CAVE	**0.8929**	**0.8948**	**0.8948**	0.9801

Table 7. Overall ablation results at $T = 5$ on FaceForensics++ (c23). Best values are shown in bold.

	Original		Face2Face		FaceShifter		FaceSwap	
Model Variant	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
Backbone only	**0.9214**	0.8062	0.8357	0.8966	0.8714	0.8905	0.8786	**0.9283**
Backbone + TRAP	0.8786	0.7987	0.8500	0.8947	0.9143	0.9046	0.8643	0.9202
Backbone + CAVE	0.7929	0.7475	0.8143	0.8736	0.8429	0.8582	**0.9357**	0.9129
Backbone + TRAP + CAVE	0.8857	**0.8212**	**0.9000**	**0.9333**	**0.9214**	**0.9149**	0.8645	0.9098

Table 8. Per-class Accuracy and F1 for the $T = 5$ ablation. Best value in each column is shown in bold.

model is conditioned on detector-selected evidence rather than on an independently verified causal signal. For this reason, the generated explanations should be interpreted as detector-guided rather than as guaranteed faithful accounts of the detector's internal reasoning. Developing more robust evidence-to-ROI linking and studying how to reduce detector bias in the explanation stage remain important directions for future work.

Additional Reproducibility Details for the Qwen2.5-VL Explainer

The Qwen2.5-VL explainer is trained with DeepSpeed ZeRO Stage 2 on four NVIDIA RTX A6000 GPUs, each equipped with 48 GB of memory. Both optimizer offloading and parameter offloading are disabled, whereas communication overlap is enabled to improve distributed training efficiency. We disable `torch.compile` because it is incompatible with dynamic vision patches in our implementation.

The model is fine-tuned for 5 epochs with a batch size of 1 per GPU and gradient accumulation of 8, yielding an effective batch size of 32 across 4 GPUs. We use AdamW with the default Hugging Face implementation and a cosine learning-rate schedule. The learning rate is 2×10^{-5} for $t \in \{1, 5\}$ and 1×10^{-5} for $t = 3$, with a warmup ratio of 0.03 and weight decay of 0.01. The maximum number of pixels per image is limited to 300,000, and 8 DataLoader workers are used. We select the best checkpoint according to `eval_loss`, where lower values indicate better validation performance.

Additional Qualitative Results for Temporal Qwen2.5-VL Explanations and FLARE

True\Pred	Orig.	F2F	FSH	FS
Orig.	129	2	9	0
F2F	21	117	2	0
FSH	14	2	122	2
FS	16	0	1	123

(a) Backbone only

True\Pred	Orig.	F2F	FSH	FS
Orig.	123	6	10	1
F2F	19	119	2	0
FSH	10	1	128	1
FS	16	0	3	121

(b) Backbone + TRAP

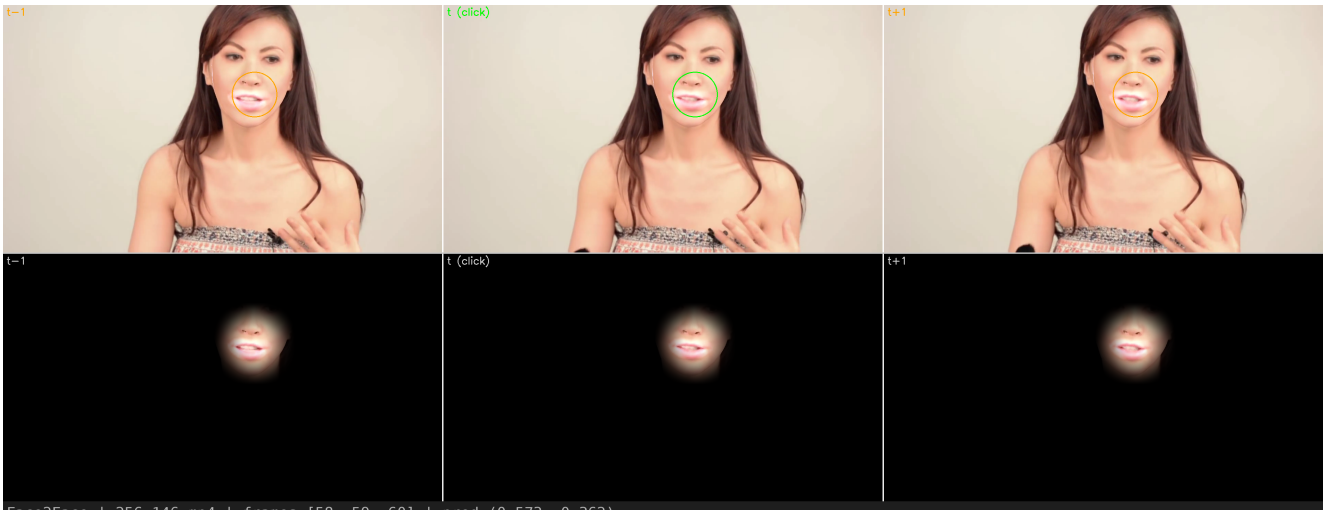
True\Pred	Orig.	F2F	FSH	FS
Orig.	111	6	11	12
F2F	21	114	3	2
FSH	19	1	118	2
FS	6	0	3	131

(c) Backbone + CAVE

True\Pred	Orig.	F2F	FSH	FS
Orig.	124	4	8	4
F2F	12	126	2	0
FSH	10	0	129	1
FS	16	0	3	121

(d) Backbone + TRAP + CAVE

Table 9. Confusion matrices for the $T = 5$ ablation study. Orig.: Original, F2F: Face2Face, FSH: FaceShifter, FS: FaceSwap.



Face2Face | 256_146.mp4 | frames [58, 59, 60] | pred (0.573, 0.362)

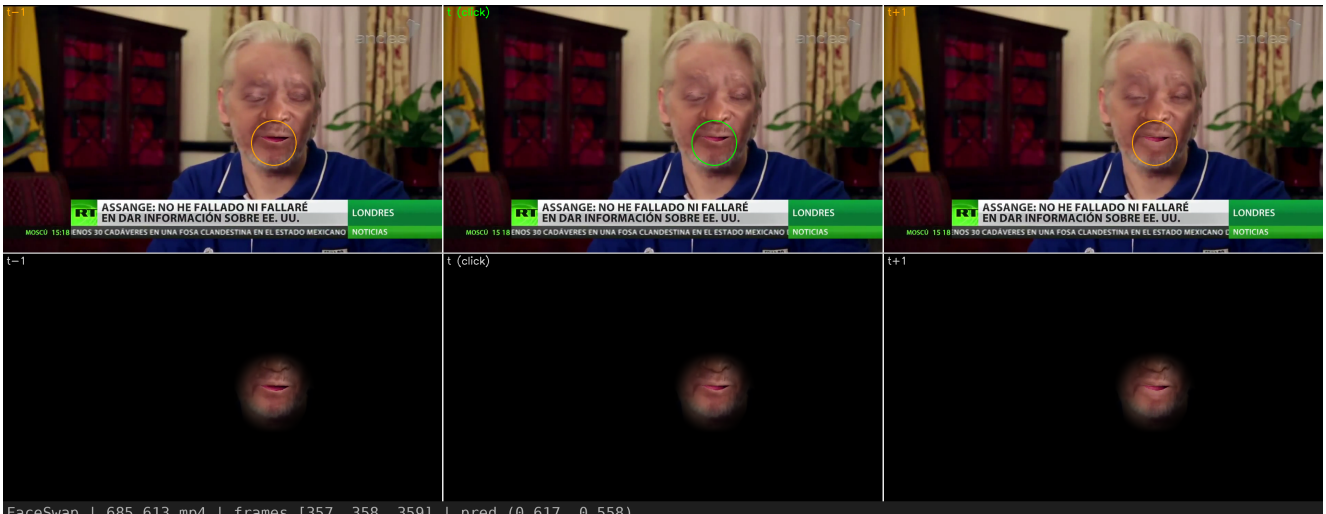
Row 1: original (green=pred) Row 2: spotlight input

MODEL:

Artifacts on the mouth and on the right side of the face.

GROUND TRUTH:

A white spot can be seen around the mouth.



FaceSwap | 685_613.mp4 | frames [357, 358, 359] | pred (0.617, 0.558)

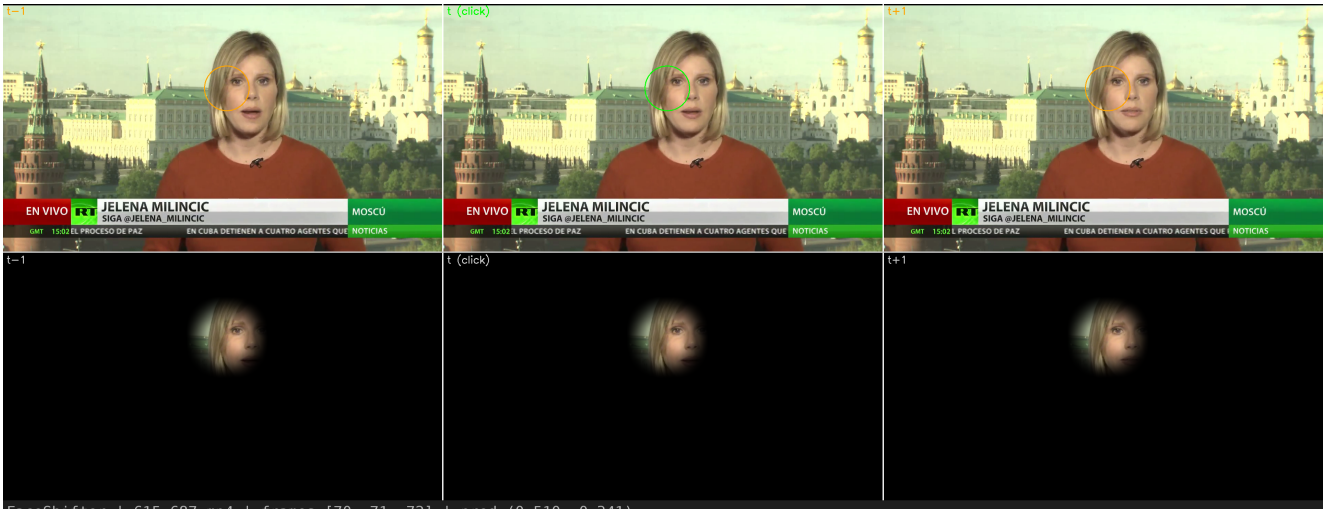
Row 1: original (green=pred) Row 2: spotlight input

MODEL:

Artifacts on the mouth and on the right side of the face.

GROUND TRUTH:

The faces looks slightly deformed and has an unusual color. You can see lingering artifacts all over the face



FaceShifter | 615_687.mp4 | frames [70, 71, 72] | pred (0.510, 0.341)

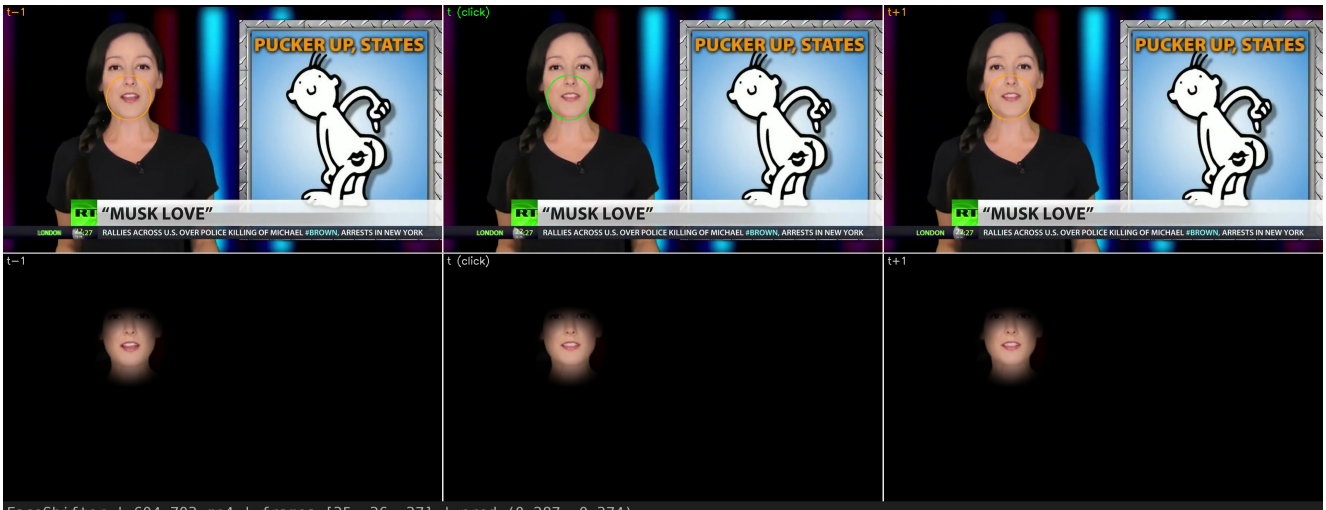
Row 1: original (green=pred) Row 2: spotlight input

MODEL:

The light is flickering on the face. Artifacts around the eyes.

GROUND TRUTH:

The strand of hair on the right is fluctuating. The eyebrows residuals are flickering.



FaceShifter | 604_703.mp4 | frames [35, 36, 37] | pred (0.287, 0.374)

Row 1: original (green=pred) Row 2: spotlight input

MODEL:

Artifacts on the mouth and on the right side of the face.

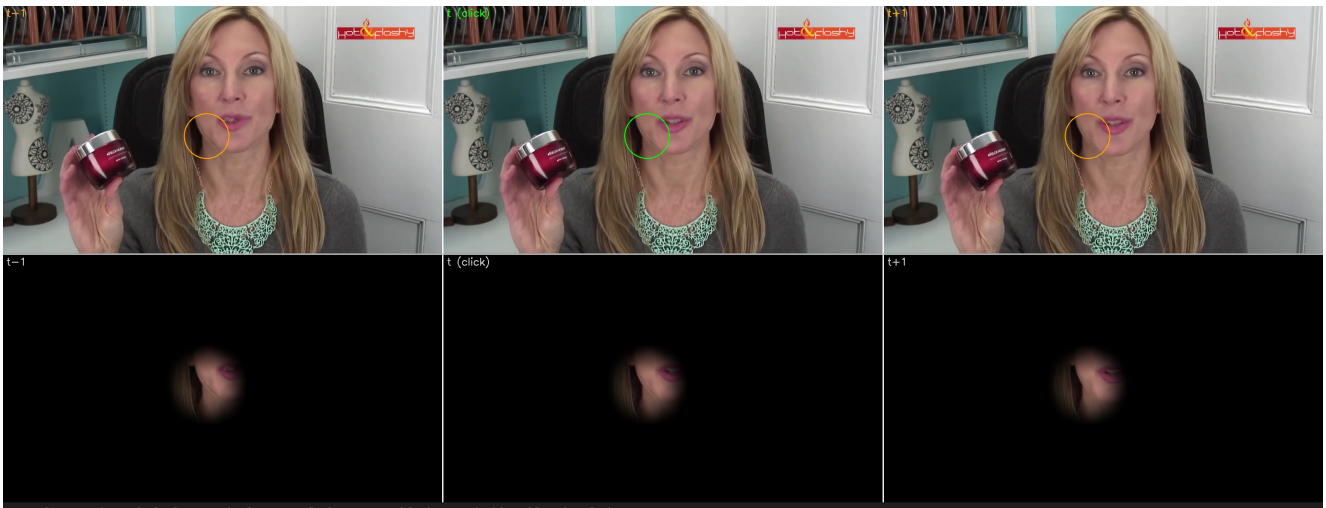
GROUND TRUTH:

The mouth is very distorted. The right ear is blurred.

Face2Face | 322_315.mp4 | frames [155, 156, 157] | pred (0.698, 0.448)
Row 1: original (green=pred) Row 2: spotlight input

MODEL:
Artifacts on the mouth and on the chin.

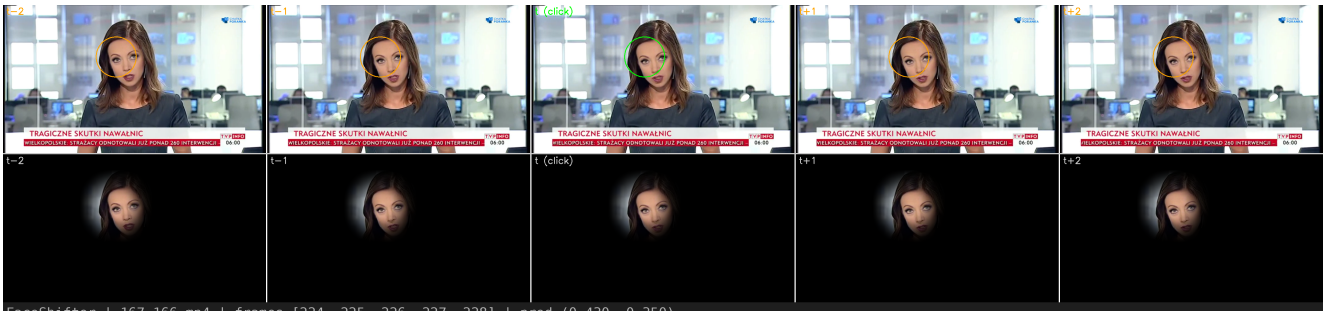
GROUND TRUTH:
Glitches on the face.



Face2Face | 146_256.mp4 | frames [16, 17, 18] | pred (0.463, 0.525)
Row 1: original (green=pred) Row 2: spotlight input

MODEL:
Artifacts on the mouth and on the chin.

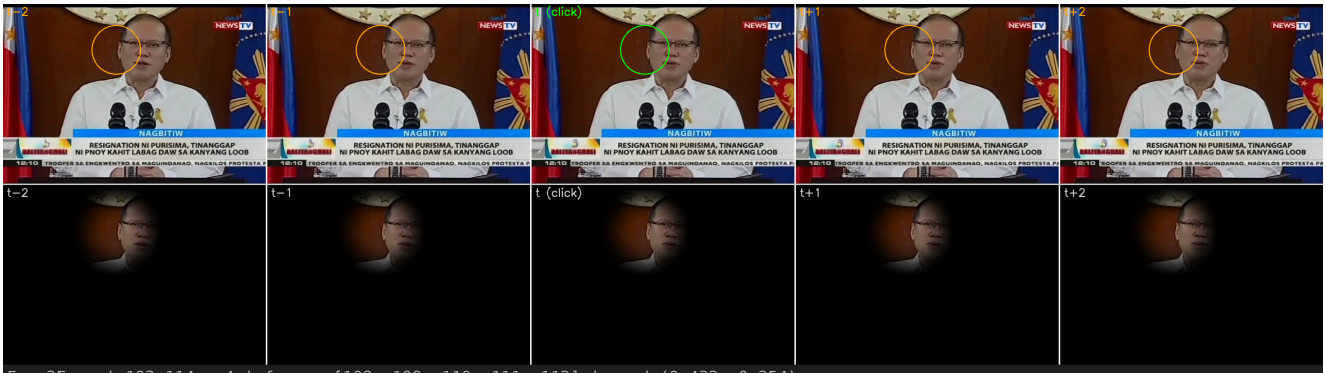
GROUND TRUTH:
Artifacts on the mouth and around the chin.



FaceShifter | 167_166.mp4 | frames [224, 225, 226, 227, 228] | pred (0.430, 0.350)
Row 1: original (green=pred) Row 2: spotlight input

MODEL:
Distortions on the face. The gaze is unnatural.

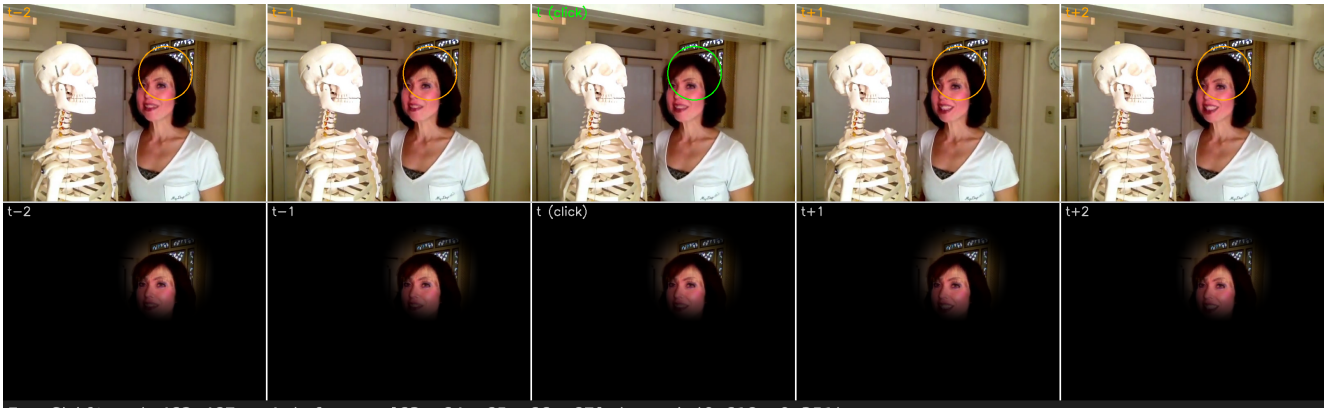
GROUND TRUTH:
Some artifacts around the eyes.



Face2Face | 102_114.mp4 | frames [108, 109, 110, 111, 112] | pred (0.432, 0.254)
Row 1: original (green=pred) Row 2: spotlight input

MODEL:
The eyeglasses have missing parts. The face is blurry.

GROUND TRUTH:
Eyeglasses seem unfinished. The mouth movement seems a bit off.



FaceShifter | 462_467.mp4 | frames [63, 64, 65, 66, 67] | pred (0.618, 0.351)
Row 1: original (green=pred) Row 2: spotlight input

MODEL:
Artifacts on the forehead. Eyes have a strange light.

GROUND TRUTH:
Unnatural gaze. The forehead looks strange with glitches on it.

Face2Face | 872_873.mp4 | frames [64, 65, 66, 67, 68] | pred (0.502, 0.436)
Row 1: original (green=pred) Row 2: spotlight input

MODEL:
Artifacts on the mouth and on the chin.

GROUND TRUTH:
The mouth is not moving naturally when talking.

Face2Face | 563_537.mp4 | frames [207, 208, 209, 210, 211] | pred (0.522, 0.513)

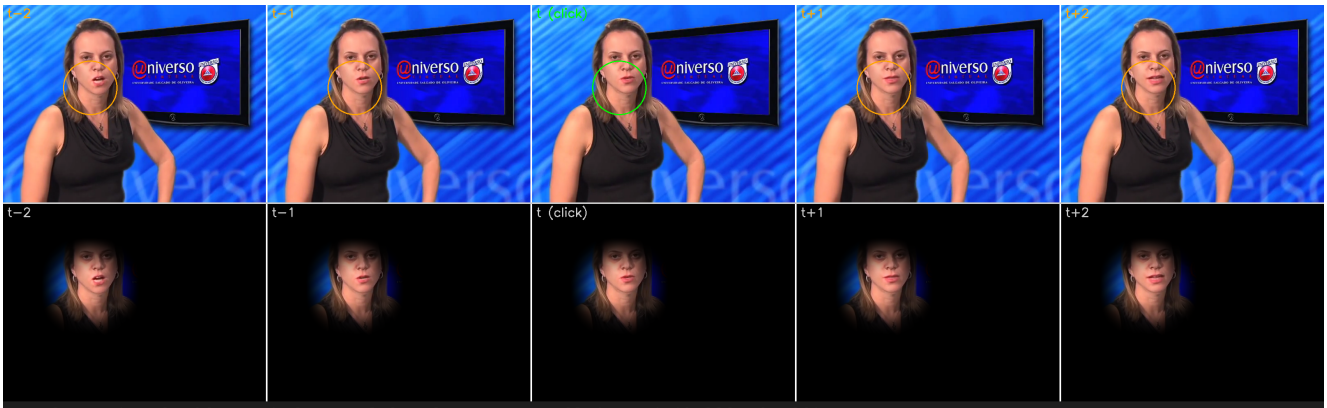
Row 1: original (green=pred) Row 2: spotlight input

MODEL:

The mouth is blurred and experiences some distortions.

GROUND TRUTH:

There are some small noisy artifacts on the nostrils. The mouth is blurred and has some visual anomalies.



Face2Face | 478_506.mp4 | frames [42, 43, 44, 45, 46] | pred (0.332, 0.420)

Row 1: original (green=pred) Row 2: spotlight input

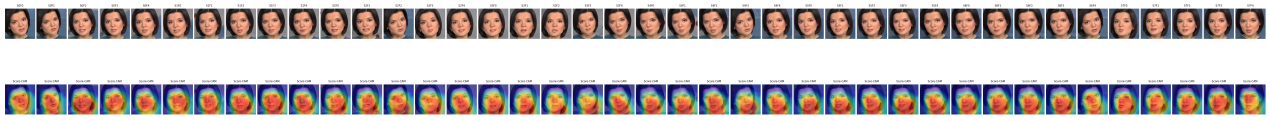
MODEL:

Artifacts on the chin. Mouth has distortions.

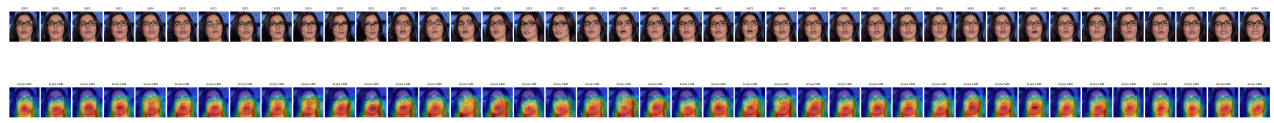
GROUND TRUTH:

Face has overall an unnatural look and the mouth has some distortions.

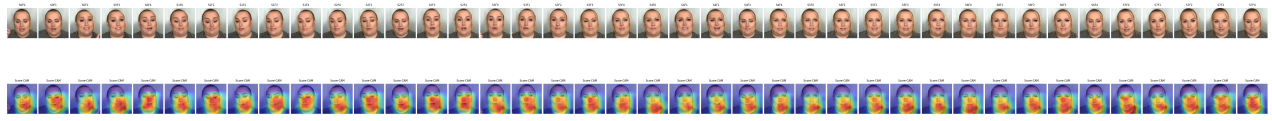
[CORRECT] True: Face2Face | Predicted: Face2Face (99.3%)
Video: 383_376.mp4 | DILD-FLARE Interpretability | T=5



[WRONG] True: FaceShifter | Predicted: Original (50.0%)
Video: 227_169.mp4



[CORRECT] True: FaceShifter | Predicted: FaceShifter (99.1%)
Video: 507_418.mp4



[WRONG] True: Face2Face | Predicted: Original (68.9%)
Video: 190_176.mp4