ISSN: 2664-9640

The Surface Consistency Trap: Evaluating LLM-Generated Deepfake Explanations

Hubin Wang*

School of Physics and Electronic Engineering, Jiangsu University, Zhenjiang 212013, China

*Corresponding author: Hubin Wang (Email w820914l@gmail.com)

Abstract

Automatic metrics can reward surface agreement while missing factual and logical errors. We define the Surface Consistency Trap as the systematic overestimation of explanation quality when surface-level signal consistency is treated as evidence of deep logical correctness. We introduce the Dual-Layer CFG framework for evaluating LLM-generated deepfake explanations from structured forensic inputs. Layer 1 (Auto-CFG) checks surface consistency, whereas Layer 2 (Human-CFG) assesses D1 (Numerical Correctness), D2 (Threshold Judgment), and D3 (Causal Reasoning). In a retrospective audit of 5,000 legacy explanations, the mean graded CSF coverage score is 93.5%. On the matched N = 101 subset, the binary Auto-CFG signal-consistency pass rate is 91.1%, while both human raters assign a strict Human-CFG pass rate of 0.0% (95% CI, 0.0–3.7%). A separate API-based GLM-4 cross-system validation (N = 300) reaches 99.7% on D1 and D2 but only 1.0% on D3 and strict pass. A post-repair audit (N = 300) uses validated visual TFL evidence with a constrained local FLAN-T5-base+LoRA generator; it reaches 100.0% on D1/D2 and D3=n for all items because the schema explicitly abstains from unsupported mechanism claims. A human baseline (N = 45) achieves 100% strict pass. These distinct arms show that input validity, surface fidelity, and mechanism-level reasoning require separate evaluation.

Keywords

Deepfake Detection; Large Language Models; Faithfulness Evaluation; Numerical Hallucination; Surface Consistency Trap.

1. Introduction

Deepfake detection has evolved from pure classification tasks to explainable systems where users demand transparency in forensic decisions [1]. With the rise of Multimodal Large Language Models (MLLMs), researchers have begun generating natural language explanations for detected forged regions, enabling non-expert users to understand why a video is flagged as manipulated [2-4]. However, a critical question remains largely unexamined: to what extent do these generated explanations accurately reflect the underlying forensic evidence?

Existing automatic evaluation metrics primarily check whether detected signals are mentioned in the generated rationale. While these metrics frequently achieve high scores (e.g., 90%+), they fail to detect three critical failure modes: (1) numerical errors-the model claims "spectral metrics (1004.94)" when the actual value is 10.04; (2) template repetition-the model reuses the same sentence structure regardless of signal strength; and (3) missing causal reasoning-the model states what is present but never explains why it indicates forgery.

We use the term Surface Consistency Trap for cases in which a surface metric yields a high pass rate while a stricter human-correctness metric yields a substantially lower rate on the same or explicitly contrasted evaluation samples. In the matched Signal-Guided Constraint Generation

ISSN: 2664-9640

(SGCG) audit subset, the binary Auto-CFG signal-consistency pass rate (91.1%) and strict Human-CFG (0.0%) differ by 91.1 percentage points.

Through retrospective evaluation of 5,000 explanations produced under the legacy structured-input configuration and human verification of 101 stratified samples, we observe a 93.5% full-batch mean graded CSF coverage score. In the matched audit subset, the binary Auto-CFG signal-consistency pass rate is 91.1%, whereas strict Human-CFG is 0/101 for both raters (0.0%, 95% CI [0.0%, 3.7%]). In a separate API-based GLM-4 cross-system arm (N = 300), D1 and D2 each reach 99.7%, whereas D3 and strict pass remain 1.0% (3/300, 95% CI [0.3%, 2.9%]). A distinct post-repair Human-CFG audit (N = 300) evaluates a local FLAN-T5-base+LoRA generator using visual-only TFL inputs. A human baseline (N = 45) achieves 100% strict pass (45/45; 95% exact CI [92.1%, 100.0%]).

Our contributions are: (1) we demonstrate and quantify the Surface Consistency Trap in retrospective deepfake-explanation audits; (2) we propose a Dual-Layer CFG framework separating surface consistency (Auto-CFG) from deep logical correctness (Human-CFG) with dimensions D1–D3; (3) we identify a 100× Numerical Hallucination subset (25.7%, 26/101) and high rater-specific Template Repetition rates (90.1–99.0%); (4) we test the D3 bottleneck in an independent API-based GLM-4 cross-system arm (N = 300); and (5) we repair the evidence pipeline, validate a visual TFL ensemble, and separately audit the resulting constrained FLAN-T5-base+LoRA explanations.

2. Related Work

Deepfake detection and explanation traditionally focus on binary classification accuracy (AP, AUC). Recent works explore MLLM-based explanations: X^2-DFD [2] evaluates forgery-feature capabilities and strengthens weak features; FakeShield [3] combines image-forgery detection, localization, and textual justification; and DD-VQA [4] grounds detection in commonsense reasoning. Recent work also uses an MLLM as a judge of deepfake reasoning quality [5]. To our knowledge, these studies do not operationalize the specific combination of numerical correctness, forensic threshold judgment, and evidence-specific causal reasoning used in Human-CFG.

Faithfulness evaluation in NLP shows that standard automatic metrics can miss factual errors identified by human raters [6]. Reviews of explainable NLP further distinguish human-authored explanations from model-generated rationales and emphasize task-specific evaluation design [7]. The deepfake setting adds numerical precision, forensic threshold judgment, and multimodal signal grounding. Numerical hallucination in LLMs has been linked to tokenization choices that fragment decimal numbers into arbitrary chunks [8,9], providing a plausible, but not experimentally isolated, explanation for the observed 100× amplification pattern.

3. Dual-Layer CFG Framework

3.1. Auto-CFG (Layer 1): Surface Consistency

In the retrospective legacy audit, Auto-CFG measures six surface-level checks: (1) JSON validity; (2) completeness (all required fields: time_range, verdict, rationale); (3) time consistency (±2s); (4) signal consistency (mentions all three recorded signal types: boundary, temporal, spectral); (5) verdict accuracy; (6) rationale length (>20 chars). Core Signal Faithfulness (CSF) is the fraction of required recorded signal types mentioned in each explanation, averaged across explanations; it is a graded coverage score rather than a binary pass rate. The legacy batch reaches a mean CSF of 93.5% (N = 5,000). The separate binary signal-consistency check passes 4,587/5,000 explanations (91.7%) and 92/101 explanations (91.1%) in the matched subset because these metrics check mentioning signals, not understanding them. A rationale can

ISSN: 2664-9640

satisfy all surface checks while containing fundamental logical errors. The audited system encoded recorded signals into a deterministic template for the FLAN-T5 generator (Listing 1); it did not receive raw video frames. This design enables a controlled behavior audit but does not validate the legacy signals as forensic evidence. Figure 1 summarizes the hierarchical relationship between the six Auto-CFG checks and the three Human-CFG dimensions.

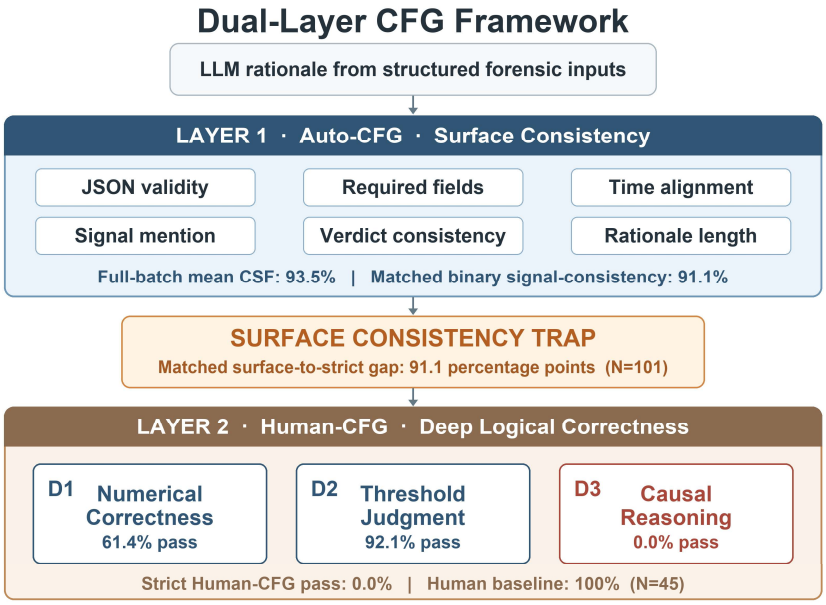


Figure 1. Dual-Layer CFG Framework. Layer 1 (Auto-CFG) checks six surface criteria, whereas Layer 2 (Human-CFG) checks D1 Numerical Correctness, D2 Threshold Judgment, and D3 Causal Reasoning. In the retrospective legacy audit, the full-batch mean graded CSF coverage score is 93.5% (N = 5,000). On the matched N = 101 subset, the binary Auto-CFG signal-consistency pass rate is 91.1%. The Human-CFG dimension rates shown use Rater A's labels; strict Human-CFG is 0.0% for both raters, a matched gap of 91.1 percentage points.

Listing 1. Prompt Template for T5 Generation

Video segment {video_id}, time range {start}s-{end}s, region: face.

Boundary artifacts: {present/absent} ({severity}), value={value}.

Temporal inconsistency: {present/absent} ({severity}), value={value}.

Texture spectral: {present/absent} ({severity}), value={value}.

Confidence: {confidence}. Generate structured verdict:

3.2. Human-CFG (Layer 2): Deep Logical Correctness

To address the Trap, we propose Human-CFG with three evaluation dimensions grounded in the logic of explanatory reasoning (Table 1).

Table 1. Human-CFG Evaluation Dimensions

Dim	Question	Pass (y)	Partial (p)	Fail (n)
D1	Do rationale values match signal values?	Exact match (±0.1)	Approximate (±1.0)	Major error (e.g., 100×)
D2	Does rationale make explicit threshold judgments?	"exceed" or "above threshold"	Implicit comparison	No threshold judgment
D3	Does rationale explain why signals indicate forgery?	Specific mechanism	Template or generic phrase	Template filler or none

ISSN: 2664-9640

Strict: D1=D2=D3=y. Relaxed: D1=y, D2=y, and D3∈{y,p}. For matched samples, the Auto-CFG–strict Human-CFG difference is the primary quantitative contrast.

Failure Mode 1: Numerical Hallucination. The 100× amplification subset occurs in 26/101 explanations (25.7%; Table 2); the 34/101 D1 failures in Table 4 also include other major numerical errors. The repeated decimal-shift pattern may reflect tokenization or decoding-format effects rather than random variation, but this study does not isolate the underlying mechanism [8,9].

Table 2. Numerical Hallucination Examples

Video	Signal Value	Rationale Value	Error Pattern
005651	S=25.92	2521.92	100× amplification (×97.2)
011980	S=10.04	1004.94	100× amplification (×100.1)
034053	S=9.42	942.42	100× amplification (×100.0)
057628	S=12.01	1201.01	100× amplification (×100.0)

Failure Mode 2: Template Repetition (Rater A: 99.0%, 100/101; Rater B: 90.1%, 91/101).

Operational definition. An explanation is coded as template-repetitive when it uses the same sentence structure as more than ten other explanations in the 101-case audit, regardless of its D3 label. Under this rule, Rater A labeled 100/101 explanations (99.0%) and Rater B labeled 91/101 (90.1%) as template-repetitive; the raters agreed on 92/101 labels. This label is analyzed separately from D3: generic wording often coincides with D3=partial, but repetition alone does not determine causal correctness. The recurring sentence-'edge density and frame difference exceed normal ranges, while spectral metrics confirm synthetic artifacts'-appears across signal strengths, indicating surface pattern matching rather than evidence-adaptive reasoning.

3.3. Signal Extraction and Evidence-Schema Repair

Legacy detection-stage preprocessing and temporal localization. The retrospective SGCG (N = 5,000; Human-CFG N = 101) and GLM-4 (N = 300) arms used the supplied LAV-DF dev-set pipeline [10]. Following clip-based spatiotemporal feature extraction [11], the implementation operationalized this stage with a Kinetics-400-pretrained R3D-18 backbone [12]. RGB frames were resized to 112×112, scaled to 0–1, and normalized with channel means (0.43216, 0.394666, 0.37645) and standard deviations (0.22803, 0.22145, 0.216989). Each 16-frame window (stride 8) produced a 512-dimensional feature for a two-layer temporal head. This historical stage retained at least one interval for 4,488 of 15,521 visually modified dev videos (28.9%); its outputs define the legacy behavior-audit cohorts, not the repaired evidence pipeline.

Legacy B/T/S audit and retirement. The retrospective explanations received three pixel-derived fields (Boundary, Temporal, and Spectral). Audit found that the stored Spectral decision had been inherited from a different frequency–photometry statistic and transferred to a non-commensurate block-DCT feature. We therefore recalculated pixel evidence under one frozen pixel-v3.0 schema. The same-schema high-AC spectral ratio achieved dev AUROC 0.6094, balanced accuracy 0.5880, and an audio-only negative-control false-positive rate of 0.632; Boundary, Temporal, Spectral, and their combination all failed the prespecified development gate. Pixel B/T/S values are consequently retired from forensic claims and explanation training. They remain only as historical recorded inputs for auditing model behavior; D2 in the legacy cohorts measures consistency with those recorded inputs, not forensic validity. No post hoc threshold or severity bands are assigned to the retired B/T/S fields because doing so would imply unsupported forensic validity; the replacement TFL score is interpreted only relative to the frozen threshold specified below.

ISSN: 2664-9640

Schema-locked replacement evidence. Temporal forgery localization (TFL) uses pre-extracted 512-dimensional visual spatiotemporal features from the Kinetics-400-pretrained R3D-18 backbone [12]. Its training manifest contains 40,000 records-10,000 each for real, audio-only, video-only, and audio+video-with real and audio-only as visual negatives. Three models were trained for 10 epochs with seeds 42/43/44 and combined by an unweighted mean. Dev-A was used to lock the model and interval_mean_score threshold (0.0447634266068537). On untouched held-out Dev-B (6,000 rows; 2,000 visual pairs), the locked ensemble reaches AUROC 0.9726, average precision 0.9758, sensitivity 0.9075, specificity 0.9165, and balanced accuracy 0.9120 (95% bootstrap CI [0.9033, 0.9200]). The audio-only negative-control false-positive rate is 0.308 (95% CI [0.279, 0.337]), which passes the prespecified research gate but is not deployment-level robustness. On a source-isolated 3,200-record LAV-DF explanation pool, AUROC is 0.9383 and balanced accuracy is 0.8500. TFL is visual-only and cannot identify audio manipulation or establish a specific forgery mechanism.

4. Experiments

Dataset and cohort separation. The legacy behavior audit comprises 5,000 SGCG explanations from 4,488 LAV-DF videos, the matched 101-case Human-CFG subset (34/34/33 descriptive bins), and a separate 300-case API-based GLM-4 cross-system sample. The repaired evidence path is evaluated on locked Dev-B and then on a source-isolated 3,200-record explanation pool (1,284 sources). From that repaired pool, the post-repair Human-CFG audit selects 300 distinct source IDs: 75 each from real, audio-only, video-only, and audio+video, with 25 near-, 25 medium-, and 25 far-threshold-margin cases within each stratum. Training, validation, and explanation-test source IDs are disjoint. Dev-B and the explanation pool are held-out within LAV-DF, not external cross-dataset tests.

Statistical analysis. Wilson score 95% confidence intervals are used for binomial pass rates. For the boundary-case human baseline (45/45), a two-sided Clopper–Pearson exact interval is used (95% CI [92.1%, 100.0%]). The repaired TFL evidence evaluation uses a 2,000-repetition paired cluster bootstrap for visual pairs and row bootstrap for the audio-only negative control.

4.1. Retrospective Auto-CFG Results (N = 5,000)

Table 3. Retrospective Auto-CFG Metrics (N = 5,000)

Metric	Value
Mean graded CSF coverage	93.5%
Binary Signal-Consistency Pass	91.7%
All Checks Passed	84.2%

As shown in Table 3, the legacy mean graded CSF coverage score is 93.5%, the binary signal-consistency pass rate is 91.7%, and All Checks Passed is 84.2%. These rates are behavior-audit outcomes under the retired structured-input configuration, not estimates of forensic evidence validity or deep correctness. On the matched N = 101 subset, the binary signal-consistency pass rate is 91.1%, whereas strict Human-CFG is 0.0% for both raters.

4.2. Retrospective Human-CFG Results (N = 101)

Table 4. Human-CFG Scores on Recorded Legacy Inputs (N = 101, Rater A)

Dimension	Pass	95% CI	Partial	Fail
D1: Numerical Correctness	61.4% (62)	[51.6, 70.3]	5.0% (5)	33.7% (34)
D2: Threshold Judgment	92.1% (93)	[85.1, 95.9]	0.0% (0)	7.9% (8)
D3: Causal Reasoning	0.0% (0)	[0.0, 3.7]	92.1% (93)	7.9% (8)

ISSN: 2664-9640

As shown in Table 4, Strict pass (D1=D2=D3=y): 0.0% (0/101, 95% CI [0.0%, 3.7%]). Table 4 reports Rater A consistently: D1 has 34/101 major failures, of which 26/101 are the separately coded 100× amplification subset; D2 has 93/101 passes; and D3 has 93/101 partial and 8/101 fail labels. Rater B independently records D1/D2/D3 pass counts of 63/92/0, with the same 0/101 strict result. Template repetition is reported separately for each rater and is not used to assign D3 by itself. The strict rate is driven entirely by D3. Because pixel B/T/S is retired from forensic claims, D2 is an input-consistency score and does not validate the recorded thresholds.

Relaxed pass (D1=y, D2=y, D3∈{y,p}): Rater A records 61.4% (62/101) and Rater B records 60.4% (61/101); both record 0.0% strict pass. The strict-to-relaxed gap therefore remains a D3 bottleneck under either rater's labels.

Human baseline: Three writers (CS graduate students with ML background) given identical structured signal data for 15 stratified videos wrote explanations. A blind fourth rater scored all 45 explanations using identical Human-CFG criteria. The human baseline achieved 100% strict pass (45/45; 95% exact CI [92.1%, 100.0%]), confirming the criteria are achievable by humans under matched input conditions. For the N = 101 audit, two independent raters achieved Cohen's κ values of 0.899, 0.936, and 1.000 for D1, D2, and D3, respectively. We therefore report Rater A counts in Table 4 and Rater B as an independent reliability check rather than constructing an undocumented hybrid consensus.

4.3. API-Based Cross-System Validation with GLM-4 (N = 300)

To test whether the observed D3 bottleneck extends beyond the FLAN-T5 [13,14] system, we evaluated the GLM-4 service (Zhipu AI) [15] through its API in a separate cross-system validation. Preserved run artifacts record text mode, service model identifier 'glm-4', temperature 0.7, and a 2,048-token response cap; all 300 calls completed on June 27, 2026 (local system time), and no model weights were deployed locally. We randomly sampled 300 LAV-DF videos with retained detector intervals and complete legacy structured inputs. GLM-4 generated JSON explanations from the same field definitions used for SGCG. Three independent human annotators with machine-learning backgrounds, trained on the same coding manual and blind to the study hypotheses, evaluated all 300 explanations using the trinary y/p/n scale; Table 5 reports the per-dimension 2-of-3 majority-vote consensus labels after alignment by video_id.

Table 5. API-Based GLM-4 Human-CFG Results (N = 300, 3-Annotator Consensus)

Dimension	Pass	95% CI	Partial	Fail
D1: Numerical Correctness	99.7% (299)	[98.1, 99.9]	0.3% (1)	0.0% (0)
D2: Threshold Judgment	99.7% (299)	[98.1, 99.9]	0.0% (0)	0.3% (1)
D3: Causal Reasoning	1.0% (3)	[0.3, 2.9]	99.0% (297)	0.0% (0)

As shown in Table 5, GLM-4 strict pass is 1.0% (3/300, 95% CI [0.3%, 2.9%]) and relaxed pass is 99.3% (298/300). D1 and D2 each reach 99.7%, whereas D3 reaches 1.0% pass and 99.0% partial. Thus near-ceiling numerical and threshold agreement does not translate into frequent evidence-specific causal explanations.

Within the shared legacy structured-input task, GLM-4 raises D1/D2 from 61.4%/92.1% in the SGCG Rater A results to 99.7%/99.7%, while D3 increases only from 0.0% to 1.0%. The binary Auto-CFG signal-consistency pass rate is 91.0% (273/300) in the GLM-4 arm, yielding a 90.0 pp gap from strict Human-CFG; the corresponding matched SGCG gap is 91.1 pp (N = 101). Because the samples are not paired, these differences are descriptive cross-system contrasts rather than causal model effects. Table 6 and Figure 2 summarize the pattern.

ISSN: 2664-9640

4.4. Descriptive Cross-System Comparison

Table 6. Descriptive Cross-System Results: SGCG Rater A (N = 101) vs. GLM-4 Majority Consensus (N = 300)

Dimension	SGCG	GLM-4	Difference	Interpretation
D1 Numerical Correctness	61.4%	99.7%	+38.3 pp	Near ceiling
D2 Threshold Judgment	92.1%	99.7%	+7.6 pp	Near ceiling
D3 Causal Reasoning	0.0%	1.0%	+1.0 pp	Still rare
Strict Pass (D1=D2=D3=y)	0.0%	1.0%	+1.0 pp	Still rare
Auto-CFG binary signal consistency	91.1%	91.0%	−0.1 pp	Nearly identical

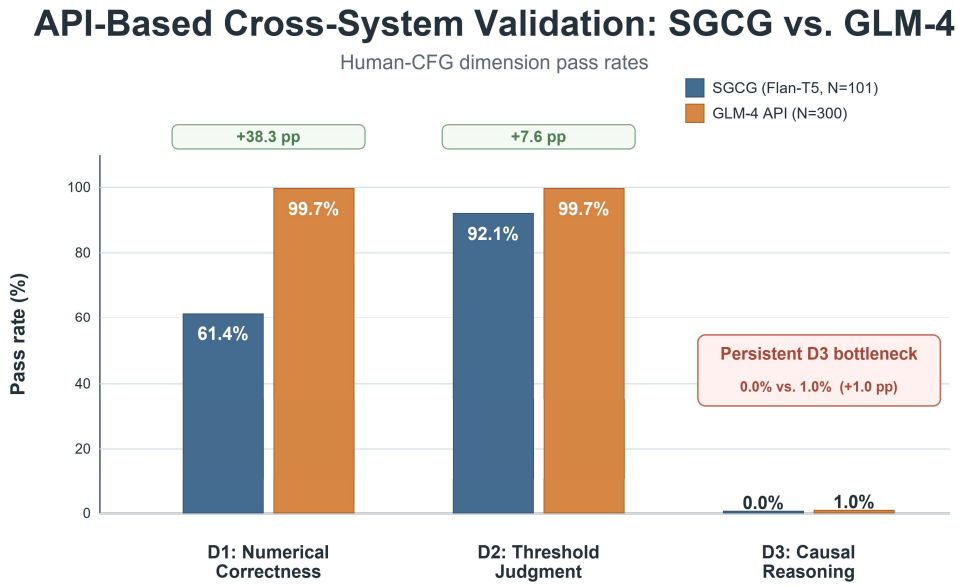


Figure 2. API-based cross-system Human-CFG comparison under the shared legacy structured-input task. The SGCG bars report Rater A (D1/D2/D3: 61.4%/92.1%/0.0%); Rater B records 62.4%/91.1%/0.0%. GLM-4 majority-vote consensus reaches 99.7% on D1 and D2, while D3 and strict pass remain 1.0%. The samples are not paired, so the bars show a descriptive cross-system pattern rather than a causal model effect.

4.5. Post-Repair TFL Schema Audit (N = 300)

Separately from GLM-4, we audited the local FLAN-T5-base model [13,14] adapted with LoRA [16] after replacing legacy B/T/S with visual TFL evidence. The model received interval_mean_score, the frozen threshold, the comparison rule, and the time interval. Its output schema was deliberately limited to above_threshold/below_threshold conclusions and an explicit no-mechanism-or-audio limitation. Three independent human annotators, trained on the same Human-CFG manual and blind to system identity and automatic results, evaluated all 300 explanations.

All 300 explanations received D1=y and D2=y; all 300 received D3=n. Strict and relaxed pass were therefore both 0/300 (strict 95% CI [0.0%, 1.3%]). The three annotators gave the same labels on all seven recorded dimensions. Auxiliary checks marked unsupported mechanism=n, overclaim=n, template repetition=n, and uncertainty calibration=y for every case. Because Human-CFG D3 requires a specific causal mechanism, D3=n in this arm records deliberate scope limitation, not contradiction of the API-based GLM-4 cross-system result. The arm demonstrates numeric/threshold fidelity and calibrated abstention; it does not demonstrate mechanism-level causal reasoning.

ISSN: 2664-9640

4.6. Retrospective Auto-CFG vs. Human-CFG Divergence

Table 7. Matched Surface-to-Deep Divergence by Evaluation Arm

Evaluation Arm	Auto Surface Pass	Human Strict Pass	Gap
SGCG matched (N=101)	91.1% (92/101)	0.0% (0/101)	91.1 pp
GLM-4 (N=300)	91.0% (273/300)	1.0% (3/300)	90.0 pp
Human baseline (N=45)	—	100.0% (45/45)	—

Note: Each row compares surface and strict Human-CFG rates within the same evaluation arm. The SGCG and GLM-4 surface endpoints use the same binary signal-consistency definition; neither validates the retired B/T/S evidence.

As summarized in Table 7, the matched SGCG arm has a 91.1 pp surface-to-strict gap and the GLM-4 arm has a 90.0 pp gap. In the N = 101 audit, 92 explanations pass the binary Auto-CFG signal-consistency check and none receives a strict pass from either human rater; thus 0/92 surface-pass cases achieve a strict pass. Because strict Human-CFG is constant at zero in this arm, item-level correlation is undefined; the percentage-point gap is the primary descriptive contrast and does not by itself establish statistical independence.

5. Discussion

Why Auto-CFG overestimates. In the legacy audit, Auto-CFG checks whether recorded fields are mentioned, not whether the evidence definition, numerical value, threshold decision, and conclusion form a valid chain. Reusable phrases such as 'exceed normal ranges' can therefore satisfy keyword checks without adapting to a specific case. The repaired pipeline removes invalid B/T/S decisions and locks the visual score, but Human-CFG remains necessary because valid input evidence alone does not create mechanism-level reasoning.

Why the D3 bottleneck is observed in both evaluated generators. D3 requires specific, evidence-adaptive mechanistic reasoning rather than accurate value copying or generic causal phrases. GLM-4 reaches 99.7% on D1/D2 but only 1.0% on D3; the three D3=y cases link case-specific evidence to a non-transferable mechanism. Together with SGCG's 0.0% D3 pass, the pattern is observed in both evaluated generators, but it does not establish model-agnostic generalization. The inference is limited to these two systems, one dataset, non-paired samples, and the retired recorded inputs. The post-repair TFL arm asks a different question: its outputs deliberately omit unsupported mechanisms, so D3=n there is expected and should be interpreted alongside the absence of overclaim and the presence of uncertainty calibration.

Comparison with NLP faithfulness evaluation. Both settings expose weak agreement between automatic and human evaluation, but deepfake explanations add three constraints: claims must be grounded in valid visual or temporal evidence, reported values must remain numerically faithful, and threshold-bearing values must be schema-matched and calibrated before supporting a conclusion. The human baseline shows that the scoring criteria are achievable under matched inputs; it does not validate the retired B/T/S fields. The repaired TFL score validates visual anomaly evidence only, not a causal forgery mechanism.

Generalizability, limitations, and conflict of interest. The API-based cross-system results are bounded by one dataset, two evaluated models, non-paired samples, and a legacy structured-input format whose pixel B/T/S fields are now retired from forensic claims. They are behavior audits, not evidence-validity estimates. The post-repair arm evaluates one local FLAN-T5-base+LoRA run under a deliberately constrained schema, so it cannot establish stability across seeds or open-ended explanation quality. The replacement TFL ensemble is visual-only, remains at 30.8% false positives on the audio-only negative control, and has not been validated on an external dataset. The evaluated SGCG system was team-developed, creating a potential

conflict of interest. Validation on FaceForensics++ [17], Celeb-DF [18], and independent published systems remains necessary. The retained API artifacts do not preserve an immutable GLM-4 model snapshot or the random-sampling seed, which limits exact rerun reproducibility.

Practical deployment trade-offs. Auto-CFG is inexpensive for rejecting invalid JSON, missing fields, or omitted evidence, but it is not a proxy for deep correctness. A defensible workflow first enforces the locked evidence schema, then runs Auto-CFG on every explanation, applies Human-CFG to a stratified audit, and escalates high-stakes cases for expert review. Legacy B/T/S outputs must never enter a forensic decision, and the visual TFL score must not be described as audio detection or mechanism proof.

Future work. Cross-dataset validation should apply the frozen TFL feature schema and held-out protocol before testing whether the D3 bottleneck persists across forgery types. Mechanism-capable explanations require additional validated evidence, less constrained generation, multi-seed robustness analysis, and counterfactual evidence flips. Human-CFG should also score calibrated abstention separately from mechanism-level success so that refusing an unsupported claim is not conflated with hallucinating one.

6. Conclusion

The Dual-Layer CFG framework separates surface compliance from deep logical correctness in LLM-generated deepfake explanations from structured forensic inputs. In the legacy SGCG audit, the full-batch mean graded CSF coverage score is 93.5%; on the matched N = 101 subset, the binary signal-consistency pass rate is 91.1% and strict Human-CFG is 0.0% for both raters. A separate API-based GLM-4 cross-system validation improves D1/D2 to 99.7% but reaches only 1.0% D3 and strict pass, while the human baseline reaches 100%. The evidence audit retires pixel B/T/S and validates a visual-only TFL ensemble. In a distinct post-repair Human-CFG arm, constrained FLAN-T5-base+LoRA outputs achieve 100.0% D1/D2 and calibrated mechanism abstention; their D3=n labels do not constitute mechanism-level success. Input validity, surface fidelity, abstention, and causal reasoning must therefore be reported as separate endpoints before high-stakes use.

References

[1] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

[2] Chen, Y., Yan, Z., Cheng, G., Zhao, K., Lyu, S., & Wu, B. (2024). X^2-DFD: A framework for explainable and extendable deepfake detection. arXiv preprint arXiv:2410.06126. https://doi.org/10.48550/arXiv.2410.06126

[3] Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., & Zhang, J. (2025). FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In International Conference on Learning Representations. https://openreview.net/forum?id=pAQzEY7M03

[4] Zhang, Y., Colman, B., Guo, X., Shahriyari, A., & Bharaj, G. (2024). Common sense reasoning for deepfake detection. arXiv preprint arXiv:2402.00126. https://doi.org/10.48550/arXiv.2402.00126

[5] Kuckreja, K., Gupta, P., Khan, M. H., & Dhall, A. (2026). Pixels don't lie (but your detector might): Bootstrapping MLLM-as-a-judge for trustworthy deepfake detection and reasoning supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 25871–25881). https://openaccess.thecvf.com/content/CVPR2026/html/Kuckreja_Pixels_Dont_Lie_But_Your_Detector_Might_Bootstrapping_MLLM-as-a-Judge_for_CVPR_2026_paper.html

[6] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906–1919). https://doi.org/10.18653/v1/2020.acl-main.173

[7] Wiegreffe, S., & Marasović, A. (2021). Teach me to explain: A review of datasets for explainable natural language processing. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 1. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/698d51a19d8a121ce581499d7b701668-Abstract-round1.html

[8] Singh, A. K., & Strouse, D. J. (2024). Tokenization counts: The impact of tokenization on arithmetic in frontier LLMs. arXiv preprint arXiv:2402.14903. https://doi.org/10.48550/arXiv.2402.14903

[9] Wallace, E., Wang, Y., Li, S., Singh, S., & Gardner, M. (2019). Do NLP models know numbers? Probing numeracy in embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5307–5315). https://doi.org/10.18653/v1/D19-1534

[10] Cai, Z., Ghosh, S., Dhall, A., Gedeon, T., Stefanov, K., & Hayat, M. (2023). Glitch in the matrix: A large-scale benchmark for content-driven audio-visual forgery detection and localization. Computer Vision and Image Understanding, 236, 103818. https://doi.org/10.1016/j.cviu.2023.103818

[11] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308). https://doi.org/10.1109/CVPR.2017.502

[12] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6450–6459). https://doi.org/10.1109/CVPR.2018.00675

[13] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67. https://jmlr.org/papers/v21/20-074.html

[14] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70), 1–53. https://www.jmlr.org/papers/v25/23-0870.html

[15] Team GLM. (2024). ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. arXiv preprint arXiv:2406.12793. https://doi.org/10.48550/arXiv.2406.12793

[16] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9

[17] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1–11). https://doi.org/10.1109/ICCV.2019.00009

[18] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3207–3216). https://doi.org/10.1109/CVPR42600.2020.00327