IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2026, accepted 10 June 2026, date of publication 22 June 2026, date of current version 26 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3706107

RESEARCH ARTICLE

GraDual: Resolving the Semantic–Frequency Trade-Off via Gradient-Regulated Dual-Branch Multimodal Manipulation Detection

DONGHUN LEE[1], JANG-HO CHOI[1], SEONGHO KIM[1], (Associate Member, IEEE), SUNGWON YI[1], AND NOJUN KWAK[2], (Senior Member, IEEE)

[1]Artificial Intelligence Safety Institute, Electronics and Telecommunications Research Institute (ETRI), Seongnam, Gyeonggi 13488, Republic of Korea
[2]Graduate School of Convergence Science and Technology, Seoul National University, Gwanak, Seoul 08826, Republic of Korea

Corresponding author: Nojun Kwak (nojunk@snu.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government (Ministry of Science and ICT, MSIT) under Grant Nos. RS-2025-02263841 and RS-2025-02215344, and in part by the Korea Internet & Security Agency (KISA) funded by the Korean Government (Personal Information Protection Commission, PIPC) under Grant Nos. RS-2026-25527246 and RS-2026-25523685.

ABSTRACT Integrating frequency-domain features into multimodal manipulation detection introduces a trade-off between localization accuracy and semantic understanding. While these features improve localization accuracy by highlighting artifacts, they degrade text grounding performance through gradient interference in unified encoders. Existing methods exhibit substantial performance degradation (15.9–17.8%) performance degradation exceeding 15–18% as manipulation complexity increases. We address this challenge through three key contributions. First, we introduce DGM4-Complex, a comprehensive benchmark of 27k image-text pairs that incorporates auxiliary manipulations to create realistic multi-level scenarios, addressing the critical limitation of existing datasets that restrict samples to single manipulations per modality. Second, we provide a systematic analysis of the semantic–frequency trade-off, revealing how frequency-domain features enhance localization accuracy while compromising text grounding performance due to gradient interference in shared encoders. Third, we present **GraDual** (Gradient-Regulated Dual-branch), a gradient-controlled framework that decouples semantic and frequency processing through differentiated gradient scaling, mitigating the trade-off and enabling robust performance across detection and grounding tasks. Our method achieves 84.78% AUC on DGM4-Complex with only a 9.1% performance decline from single-level manipulations, which is substantially lower than the 15.9-17.8% drop observed in existing methods, while maintaining stable text grounding performance (0.4% decline compared to 6.4–12.9% in others).

INDEX TERMS Deepfake, deepfake detection, generative model, multimodal.

I. INTRODUCTION

The proliferation of generative AI methods [1], [2] and advanced LLMs [3] has transformed the landscape of digital content, raising serious concerns such as deepfakes, misinformation, and fake news [4], [5], [6]. Early detection methods [7], [8] focused on single-modality manipulations. However, the recent emergence of multimodal manipulation, where both visual and textual contents are manipulated, poses new challenges for forgery detection [9], [10], [11], [12]. Multimodal manipulations maintain cross-modal consistency, which helps conceal individual artifacts. As a result, they may appear more credible and can spread more effectively than single-modal forgeries.

Beyond single-level edits, several lines of work have addressed manipulations that combine multiple operations within a single sample. Sequential facial edits [13] apply distinct generative operations in succession, hierarchical image

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

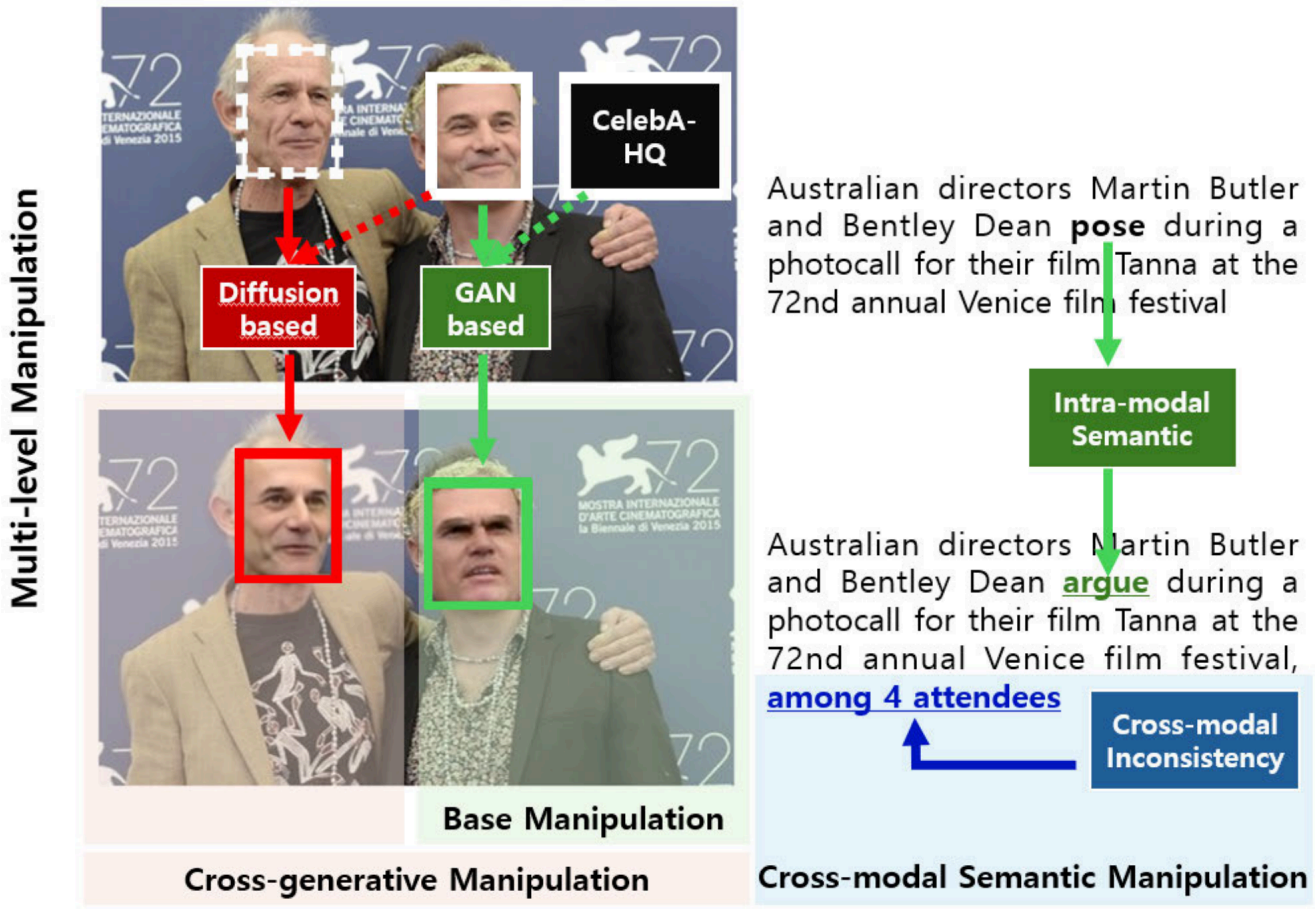


FIGURE 1. Overview of the proposed dataset DGM4-Complex. Starting from the DGM4 base, auxiliary manipulations are applied to selected samples, re-categorizing them into more complex multi-level groups. Multi-level Manipulations combine multiple synthesis methods within and across modalities to form complex manipulation patterns. Within a modality, Cross-generative Manipulations sequentially apply methods based on Generative Adversarial Networks (GANs) and diffusion models to create composite manipulation artifacts, while Cross-modal Semantic Manipulations introduce contradictions between textual descriptions and visual content to assess cross-modal reasoning.

forgeries [14] stack manipulations across semantic levels, and diffusion-based benchmarks [15] introduce composite artifacts combining multiple generation processes. Earlier forensics studies [16], [17] also demonstrated that combining manipulation types significantly degrades detector performance. Despite this evidence, existing multimodal benchmarks still restrict each sample to a single manipulation per modality, leaving multi-level robustness largely untested in the multimodal setting. To the best of our knowledge, no existing multimodal benchmark evaluates detector behavior under heterogeneous multi-level forgery patterns.

Recent multimodal detection frameworks jointly analyze image-text pairs. HAMMER [9] introduced manipulation-aware contrastive learning, and HAMMER++ [10] employs hierarchical reasoning while CSCL [11] leverages consistency learning. However, these methods primarily evaluate single-level manipulations with one manipulation type per modality, leaving their behavior under realistic multi-level forgery patterns largely uncharacterized.

Our analysis reveals a semantic-frequency trade-off in multi-level detection. While prior work [12] identified this trade-off in single-level manipulations, our findings show that this trade-off becomes more pronounced under multi-level manipulations with heterogeneous manipulation patterns. Unlike prior work that addresses this trade-off through frequency-aware architectural modules, we attribute it to gradient-level interference between competing task heads and resolve it through differentiated gradient regulation.

This trade-off is particularly evident in the proposed DGM4-Complex dataset, which introduces multi-level manipulations within and across modalities to create complex forgery patterns. An overview of the dataset construction process is shown in Figure 1.

Within single modalities, cross-generative manipulations (sequential applications of distinct generative architectures, e.g., GAN followed by diffusion) produce composite artifact distributions that challenge single-generator detection assumptions. At the cross-modal level, cross-modal semantic manipulations introduce deliberate contradictions between

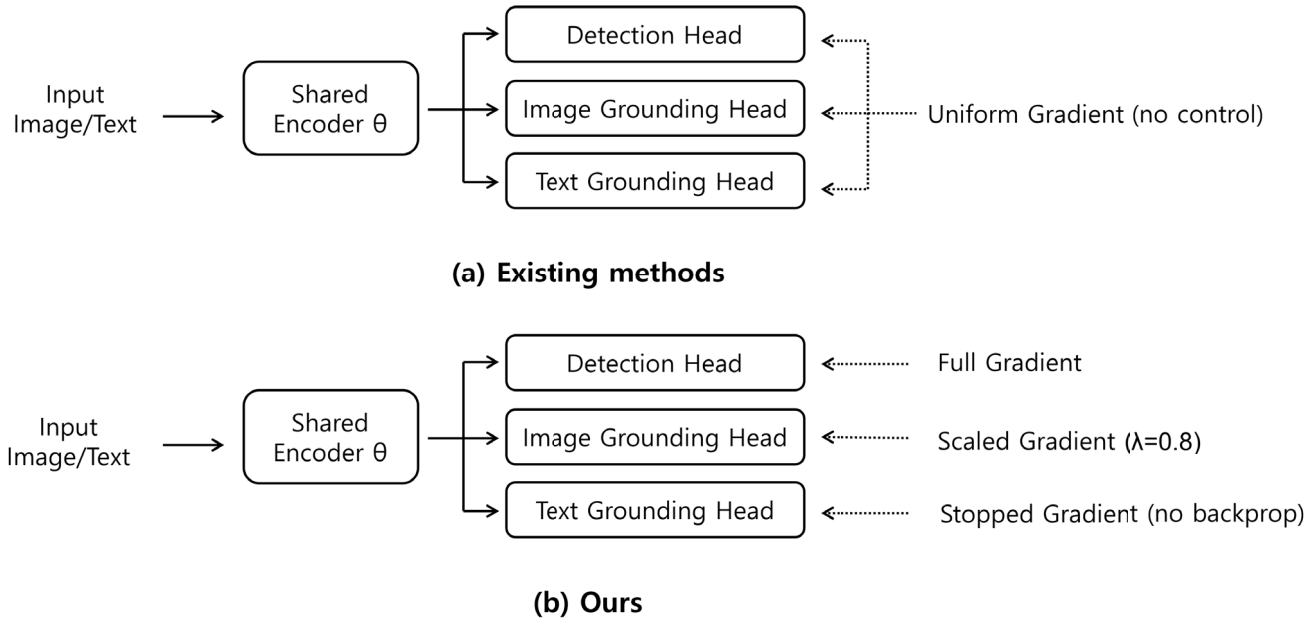


FIGURE 2. Comparison of gradient propagation between existing methods and the proposed approach. (a) Existing frameworks propagate identical gradients across tasks. (b) GraDual regulates gradient propagation for each task head, mitigating interference while preserving a shared encoder.

visual content and textual descriptions, requiring detection frameworks to verify semantic alignment beyond traditional intra-modal analysis.

The optimization dynamics of existing frameworks explain this performance drop. As illustrated in Figure 2, existing methods [9], [10], [11], [12] employ unified encoders that receive identical gradient updates from all task heads during backpropagation. This uniform optimization can introduce gradient interference. Localization objectives that prioritize frequency-domain features directly conflict with semantic coherence requirements for detection and text grounding. Gradients from localization heads push shared encoder parameters toward high-frequency artifact-sensitive representations, while gradients from detection and text grounding heads pull the same parameters toward semantically abstract features. The resulting parameter update is a compromise that satisfies neither objective optimally. This interference becomes more pronounced under multi-level manipulations, where heterogeneous manipulation types require diverse feature representations simultaneously.

To address the optimization challenges, we propose GraDual (Gradient-Regulated Dual-branch), a gradient-controlled framework with differentiated propagation strategies. The framework applies full gradients for detection to preserve semantic reasoning, scaled gradients ($\lambda = 0.8$) for localization to balance artifact sensitivity, and stopped gradients for text grounding to isolate textual optimization. By decoupling semantic and frequency processing pathways, this approach mitigates interference between competing objectives and stabilizes multi-task optimization.

Our contributions are:

- We introduce **DGM4-Complex**, a comprehensive multimodal manipulation benchmark designed to evaluate multi-level scenarios, addressing a gap not covered by prior multimodal benchmarks. The benchmark extends DGM4 with 27k image-text pairs that combine heterogeneous generation methods within and across modalities, and incorporates a dedicated control category for direct measurement of false-positive robustness, previously underexamined in multimodal forgery benchmarks.
- We provide a **systematic empirical analysis** of the semantic-frequency trade-off, attributing it to gradient interference in shared encoders rather than feature-level conflicts. Frequency integration alone degrades text grounding by 6.21% relative on DGM4-Complex, reframing the trade-off as an optimization problem rather than an architectural one addressed by prior frequency-aware methods.
- We propose **GraDual**, a gradient-regulated dual-branch framework that resolves the trade-off through differentiated gradient scaling across task heads (full gradients for detection, $\lambda = 0.8$ for image grounding, stop-gradient for text grounding) over a shared encoder. The framework achieves 84.78% AUC on DGM4-Complex with only 9.1% degradation from single-level manipulations, substantially below the 15.9–17.8% observed in prior methods, with consistent performance maintained across three independent external benchmarks.

II. RELATED WORKS

A. UNIMODAL DEEPFAKE DETECTION

Early deepfake detection relied on CNN architectures, with XceptionNet [7] and CNNDetection [8] demonstrating strong performance on GAN-generated content. Vision Transformers demonstrated improved modeling

of long-range dependencies [18], though performance varies across generative methods [19]. Frequency analysis emerged as a complementary approach, revealing distinctive spectral signatures of generative models [20], [21]. These methods identified that GANs struggle to reproduce natural high-frequency distributions with recent approaches leveraging geometric inconsistencies [22] and reconstruction errors [23], [24] for enhanced detection accuracy. F3-Net [25] mines local frequency statistics, and high-frequency features [26] improve generalization across manipulation types. UFAFormer [12] further incorporated a discrete wavelet transform for frequency-aware detection. Image forgery localization methods [17], [27], [28] and diffusion-specific detectors [29], [30], [31] provide further image-only signals. However, these unimodal approaches assume detectable single-modality artifacts, an assumption that breaks down under coordinated cross-modal manipulation.

B. MULTIMODAL MANIPULATION DETECTION

With the emergence of multimodal manipulation, recent frameworks jointly analyze image-text pairs to detect and localize manipulations. HAMMER [9] introduced hierarchical reasoning through manipulation-aware contrastive learning, while HAMMER++ [10] enhanced aggregation mechanisms for fine-grained detection. CSCL [11] leveraged contextual-semantic consistency matrices across modalities, and UFAFormer [12] integrated frequency-aware representations for joint detection and grounding. However, these frameworks evaluate only single-level manipulations and treat detection, image grounding, and text grounding as a single joint objective, leaving gradient interference between competing task heads unaddressed.

Another line of multimodal work focuses on binary detection without grounding. UniversalFakeDetect [32] employs frozen CLIP features, FatFormer [33] adapts CLIP through forgery-aware adapters, and recent CLIP-based detectors [34], [35], [36], [37] further exploit pretrained representations through lightweight adapters or frozen-backbone classifiers. A related direction targets cross-modal misinformation, where out-of-context image misuse is detected through image-text mismatch [38], external evidence [39], or multimodal large language models [40]. Similar cross-modal consistency principles underlie audio-visual deepfake detection [41], [42], [43], [44], [45], [46]. These directions inform our DGM4-Complex benchmark, which incorporates cross-modal text manipulation (TAM, Section III) to evaluate factual consistency between visual and textual content.

C. GENERALIZATION AND ROBUSTNESS

Generalization across different generation methods remains a critical challenge in deepfake detection. Representative forgery mining [47] improved domain adaptation by identifying transferable forgery prototypes across generators. Self-supervised adversarial perturbation learning [48] synthesizes hard examples to enhance robustness against unseen manipulations. Few-shot and meta-learning approaches [49], [50] enable adaptation to new generators with limited samples, and test-time adaptation methods further extend this direction by dynamically adjusting detector parameters under distribution shifts.

Recent studies also explore feature disentanglement for improved generalization, decomposing vision-language features into task-relevant components or isolating generative artifacts from real-image influence. Classical forensics studies [16], [17] demonstrated that combining manipulation types significantly degrades performance, findings further supported by recent datasets like HiFi-IFDL [14] and DiffForensics [15] that explore hierarchical manipulation structures. However, existing benchmarks primarily focus on isolated single-level manipulations, and the robustness of detectors under multi-level, sequential, and cross-modal manipulation scenarios remains insufficiently explored. Most of these efforts target distribution shifts at the generator level and do not consider robustness under heterogeneous manipulation patterns that combine multiple synthesis methods within a single sample.

D. MULTI-TASK LEARNING AND GRADIENT CONFLICT

Multi-task learning with a shared encoder frequently suffers from gradient conflict, where task-specific gradients counteract one another during backpropagation. Existing approaches address this through magnitude balancing [51], [52], gradient-direction manipulation [53], [54], or multi-objective optimization [55]. These methods operate at runtime, dynamically adjusting gradients during training without exploiting prior knowledge about which task should receive which type of gradient signal.

GraDual takes a different approach by encoding such prior knowledge directly. The semantic-frequency trade-off in Section IV-B shows that detection benefits from semantic abstraction, localization from high-frequency artifact cues, and text grounding from unperturbed text representations. GraDual reflects this trade-off through fixed differentiated scaling factors applied to each task head, avoiding the per-batch overhead of runtime gradient projection. The chosen scaling factor is empirically validated in Section V-C. Table 1 summarizes how GraDual differs from prior methods across the three design axes addressed in this work.

III. DGM4-COMPLEX DATASET

We introduce **DGM4-Complex**, designed to evaluate detection robustness under realistic and complex forgery patterns. Existing manipulation detection datasets limit each sample to a single manipulation per modality, failing to capture real-world forgery complexity. Current benchmarks evaluate isolated modifications such as face swapping or attribute changes in images, paired with semantic or sentiment alterations in text. We introduce **Multi-level Manipulations** that combine multiple synthesis methods within and across modalities to create complex forgery patterns. Within single

TABLE 1. Comparison of GraDual with prior methods across three design axes. "Multi-level" indicates evaluation on multi-level manipulation scenarios. "Frequency" indicates the use of frequency-domain features. "Text Grounding" indicates support for token-level text manipulation localization.

Method	Multi-level	Frequency	Text Grounding
F3-Net [25]		✓	
HAMMER [9]			✓
HAMMER++ [10]			✓
CSCL [11]			✓
UFAFormer [12]		✓	✓
GraDual (ours)	✓	✓	✓

modalities, **Cross-generative Manipulations** apply multiple generative architectures sequentially to produce composite manipulation artifacts. Specifically, the base images are manipulated using a GAN-based method, followed by an additional diffusion-based generation, resulting in hybrid forgeries that exhibit mixed frequency patterns. Across modalities, **Cross-modal Semantic manipulations** introduce intentional contradictions between textual descriptions and visual content, evaluating the model's ability to perform cross-modal reasoning. While the base dataset focuses on intra-modal semantic edits, DGM4-Complex extends this by creating explicit cross-modal inconsistencies to simulate more realistic multimodal forgeries. These multi-level manipulations challenge detection frameworks optimized for single-level manipulation cases, with our experiments revealing performance degradation over 15% against complex forgery patterns as shown in Table 7.

Starting from the complete DGM4 dataset [9], we introduce auxiliary manipulations to a subset of samples to construct multi-level forgery cases. Rather than replacing existing annotations, these auxiliary operations are layered on top of the original manipulations, resulting in more complex combinations within and across modalities. The extended dataset therefore contains both original DGM4 samples and newly constructed multi-level variants, as summarized in Table 4. This design preserves the original benchmark structure while enabling systematic evaluation under increased manipulation complexity.

A. AUXILIARY MANIPULATION TYPES

We introduce three auxiliary categories that combine with base DGM4 manipulations (FS, FA, TS, TA) to create multi-level manipulations with complex forgery patterns.

1) FACE ADDITIONAL MANIPULATION (FAM)

FAM introduces *cross-generative manipulations*, where sequential application of distinct generative architectures produces composite artifacts that challenge existing detection frameworks. Specifically, a GAN-based method is applied to one target and a diffusion-based method to another, resulting in composite manipulation artifacts that differ from single-generator outputs. FAM applies to multi-person images

TABLE 2. Face manipulation combinations. A' denotes attribute-modified A, C sampled from CelebA-HQ.

Type	Original	Base	+ FAM
Face Attribute (FA)	(A, B)	(A', B)	(A', A)
Face Swap (FS)	(A, B)	(C, B)	(C, A)

detected via MTCNN: for base manipulation $A \rightarrow A'$ (attribute) or $A \rightarrow C$ (swap) generated through GANs, diffusion models subsequently replace person B with identity A.

Base manipulations employ GAN-based methods for identity replacement (Face Swap, FS) using SimSwap [56] and InfoSwap [57], or attribute modification (Face Attribute, FA) through HFGI [58] and StyleCLIP [59] generation. Auxiliary manipulations utilize the diffusion-based FaceAdapter [60] method. This sequential generation framework encourages detectors to disentangle representations across heterogeneous generative mechanisms. Table 2 details the resulting manipulation combinations.

2) TEXT ADDITIONAL MANIPULATION (TAM)

TAM introduces *Cross-modal semantic manipulations* that introduce deliberate factual inconsistencies between visual and textual modalities, challenging detection frameworks to perform cross-modal verification beyond conventional forgery detection. Base text manipulations employ Named Entity Recognition for semantic replacement (Text Swap, TS) preserving person entities while altering context, or sentiment modification (Text Attribute, TA) using RoBERTa [61] classification and B-GST generation. Building upon these intra-modal semantic manipulations, TAM appends phrases with incorrect person counts to test cross-modal consistency. Given ground-truth count n_{true} from DGM4 metadata, we generate $n_{\text{fake}} = n_{\text{true}} + \delta$ where $\delta \in [1, 3]$ and append phrases such as "with n_{fake} people in attendance" encouraging detectors to reason about semantic correspondence between image content and textual descriptions.

3) NEUTRAL TAIL (NT AND NT_{CTRL})

NT appends semantically ambiguous phrases ("with updates pending", "with limited context") that weaken statements without contradiction, examining whether detectors identify subtle meaning shifts. The control variant NT_{CTRL} applies identical additions to authentic samples while preserving original labels, testing whether models incorrectly associate formatting with manipulation.

Table 3 presents representative text manipulation examples across auxiliary types and their combinations with base manipulations.

B. SAMPLE SELECTION AND REPRODUCIBILITY

Auxiliary manipulations follow a deterministic selection procedure that ensures reproducibility across the dataset. Each category applies category-specific eligibility criteria. FAM

TABLE 3. Text manipulation examples. Base manipulations are underlined, TAM in bold, and NT in italic. *NT_{CTRL} maintains authenticity labels.

Type	Base Text	Manipulated Text
TAM	Tom Hicks and his son at Stamford Bridge	Tom Hicks and his son at Stamford Bridge**, with six people in attendance**
TA + TAM	Ilham Aliyev is greeted by John Kerry	Ilham Aliyev is detained by John Kerry**, including four people**
TS + TAM	David Cameron announced tax powers	Anna Wintour arrives at a state dinner**, among four attendees**
NT	Brady Heslip set a record last year	Brady Heslip set a record last year*, with summary unavailable*
TA + NT	Acrostic for Fergie Very poetic	Acrostic for Fergie Very poorly*, with no information available*
TS + NT	Clinton leaves political meeting	Huffington and Morris switched sides*, with updates pending*
NT_{CTRL}	Kiefer in The Lost Boys	Kiefer in The Lost Boys*, with limited context**

TABLE 4. Dataset structure showing DGM4 original samples and DGM4-Complex extensions. Statistics are reported on the full benchmark before training-time deduplication. Single-level categories retain base manipulations only, whereas multi-level categories incorporate auxiliary manipulations that combine multiple techniques within a modality.

Category	Original	Added	Total (%)
Image Single-level	69,081	–	69,081 (24.95)
Image Multi-level	–	26,202	26,202 (9.46)
Text Single-level	23,349	9,886	33,235 (12.00)
Text Multi-level	–	6,092	6,092 (2.20)
Multimodal Single-level	25,949	–	25,949 (9.37)
Multimodal Multi-level	–	48,832	48,832 (17.63)
Control	53,693	13,847	67,540 (24.39)
Total	172,072	104,859	276,931 (100.00)

applies to multi-face images detected via MTCNN ($\geq$2 faces) with base manipulations from the face categories (FS or FA). Face replacement targets the unmodified face using an established diffusion-based face synthesis pipeline [60], and samples failing the pipeline's alignment check are skipped. TAM applies to captions with a named entity and an action verb. Captions that are abstract, crowd-referencing, or title-format are excluded to maintain count-based augmentation validity. NT applies as a complementary stylistic category, appending semantically ambiguous phrases that do not alter the factual meaning of the caption. NT_{CTRL} applies only to authentic samples and attaches the auxiliary label without modifying the text, providing a control for surface-pattern bias in detection.

All manipulated tokens receive binary annotations $y_{tok} \in \{0, 1\}^L$ for token-level localization, where L denotes sequence length.

C. DATASET STATISTICS

Table 4 summarizes the composition of DGM4-Complex, distinguishing original samples from auxiliary manipulations across single- and multi-level categories.

Table 5 shows auxiliary manipulation occurrence statistics. Note that individual samples may contain multiple auxiliary types, resulting in occurrence counts exceeding unique sample counts. FAM appears in 16.9% of total samples, TAM in 11.8%, NT in 8.0%, and NT_{CTRL} in 5.0%, creating heterogeneous patterns for robustness evaluation.

TABLE 5. Auxiliary manipulation statistics in DGM4-Complex. Multiple types may co-occur within individual samples.

Type	Occurrences	% of Dataset
FAM	46,931	16.9%
TAM	32,553	11.8%
NT	22,155	8.0%
NT_{CTRL}	13,847	5.0%
Total Occurrences	115,486	41.7%

D. ANNOTATION FRAMEWORK

DGM4-Complex provides three annotation levels: (1) sample-level binary labels $y_{auth} \in \{0, 1\}$ for detection, (2) bounding boxes $\{(x_i, y_i, w_i, h_i)\}_{i=1}^{K}$ for image grounding where $K \leq 2$, and (3) token-level binary masks $\{y_j\}_{j=1}^{L}$ for text grounding. This framework supports evaluation across detection and localization tasks.

IV. METHOD

We propose **GraDual** (Gradient-Regulated Dual-branch), a framework that balances semantic and frequency learning through differentiated gradient propagation, enabling stable multi-task optimization without interference. GraDual processes image-text pairs through specialized modules to handle heterogeneous manipulation types, providing binary detection as well as fine-grained spatial and textual grounding (Figure 3). As illustrated in Figure 2, existing frameworks propagate identical gradients across tasks, causing interference between semantic and frequency objectives. GraDual mitigates this by applying full gradients for detection, scaled gradients (λ=0.8) for localization, and stopped gradients for text grounding, effectively decoupling semantic and frequency learning and stabilizing multi-task optimization without additional architectural complexity.

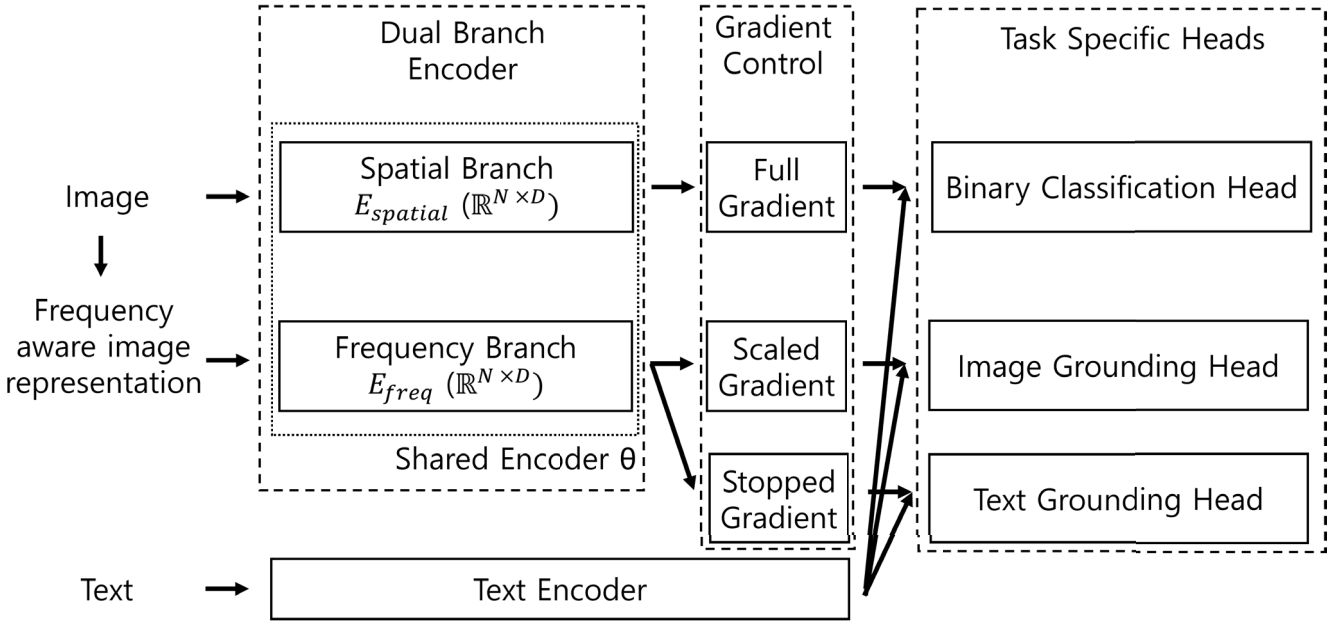


FIGURE 3. Overall architecture of the proposed **GraDual** framework. Given a multimodal input consisting of an image–text pair, the image is processed through two parallel branches: the ***Spatial Branch*** for semantic features and the ***Frequency Branch*** for frequency domain features.

A. PROBLEM FORMULATION

Given an image-text pair (I, T), our method detects and localizes manipulations across multiple levels. We define the task outputs as:

$$\mathcal{Y} = \{\mathcal{D}, \mathcal{B}, \mathcal{M}\} \subseteq \{0,1\} \times \mathbb{R}^{K\times 4} \times \{0,1\}^{L} \quad (1)$$

where $\mathcal{D} : (I, T) \rightarrow \{0, 1\}$ performs binary manipulation detection, $\mathcal{B} : I \rightarrow \{b_i\}_{i=1}^{K}$ predicts up to $K = 2$ bounding boxes parametrized as (x_i, y_i, w_i, h_i) for manipulated regions, and $\mathcal{M} : T \rightarrow \{0, 1\}^{L}$ generates token-level manipulation masks.

Our method tackles the multi-level manipulations defined in Section III, including both base manipulations (FS, FA, TS, TA) and auxiliary manipulations (FAM, TAM, NT). To address the semantic–frequency trade-off identified in Section IV-B, **GraDual** introduces two main components: (1) a dual-branch architecture that separates semantic and frequency-domain processing, and (2) a gradient regulation via scaling that balances task-specific optimization through differentiated propagation. The dual-branch design realizes two parallel input pathways through a single shared encoder, sharing all encoder parameters across branches.

B. MOTIVATION: SEMANTIC-FREQUENCY TRADE-OFF

Frequency-domain features in multimodal detection present a trade-off between localization accuracy and semantic understanding. While prior work [12] observed this on single-level manipulation detection, our analysis reveals that multi-level manipulation detection increases this conflict when detectors encounter heterogeneous manipulation types within and across modalities. Using DGM4-Complex, our comprehensive benchmark for multi-level manipulation detection, we demonstrate how complex forgery patterns intensify semantic–frequency interference.

TABLE 6. Ablation study on DGM4-Complex. Adding frequency-domain features improves localization (IoU ↑) but decreases text grounding performance (F1-tok ↓).

Method	AUC (%)	IoU (%)	F1-tok (%)
Baseline (HAMMER++)	78.41	63.12	50.86
+ Frequency features	80.53	68.96	47.70
Relative Delta	*2.70% ↑*	*9.24% ↑*	*6.21% ↓*

To investigate this trade-off, we integrate frequency-domain features into the baseline method (HAMMER++ [10]) and evaluate performance on both DGM4 and DGM4-Complex. As shown in Table 6, introducing frequency cues produces a clear trade-off: localization improves by 9.24% (relative) and detection increases by 2.70% (relative), whereas text grounding performance decreases by 6.21% (relative). This degradation, initially observed under single-level manipulations, persists in multi-level settings, indicating that semantic–frequency conflicts remain consistent across varying manipulation complexities.

C. DUAL-BRANCH FEATURE EXTRACTION

To address gradient interference, we decouple spatial and frequency-domain processing through parallel pathways. The spatial branch processes RGB images to preserve semantic information, while the frequency branch processes images transformed via Discrete Wavelet Transform (DWT) to

enhance manipulation artifacts:

$$\mathbf{E}_{\text{spatial}} = \text{Encoder}_\theta(I) \in \mathbb{R}^{N\times D} \tag{2}$$

$$\mathbf{E}_{\text{freq}} = \text{Encoder}_\theta(\text{DWT}(I)) \in \mathbb{R}^{N\times D} \tag{3}$$

where $N = HW/P^2$ denotes the number of patches with size $P = 16$, and $D = 768$ represents the feature dimension following ViT-Base architecture. The DWT employs single-level Haar wavelets, decomposing images into four sub-bands $\{LL, LH, HL, HH\}$ capturing distinct frequency characteristics. High-frequency components (LH, HL, HH) amplify manipulation traces such as splicing boundaries and compression artifacts, while low-frequency approximation (LL) provides coarse spatial structure. These sub-bands are fused through 1×1 convolution before visual encoding.

Both branches share encoder parameters θ but process different input representations. The spatial branch learns semantic features for detection and text alignment, while the frequency branch learns artifact-sensitive features for localization. This design follows multi-task learning principles without introducing additional learnable parameters in the encoder. The two branches share encoder weights and differ only in input representations. Specifically, the only difference between branches is the input transformation, and no encoder parameters are duplicated across branches. Although feature extraction is performed twice, no additional task-specific heads are introduced, resulting in limited inference-time overhead.

D. GRADIENT REGULATION VIA SCALING

To prevent gradient interference, we apply differentiated gradient scaling to each task. This approach controls how gradients from different losses update the shared encoder during backpropagation:

$$\nabla_\theta \mathcal{L}_{\text{total}} = \nabla_\theta \mathcal{L}_{\mathcal{D}} + \lambda \nabla_\theta \mathcal{L}_{\mathcal{B}} + 0 \cdot \nabla_\theta \mathcal{L}_{\mathcal{M}} \tag{4}$$

where the detection loss $\mathcal{L}_{\mathcal{D}}$ provides full gradients, the localization loss $\mathcal{L}_{\mathcal{B}}$ contributes scaled gradients ($\lambda = 0.8$), and the token grounding loss $\mathcal{L}_{\mathcal{M}}$ applies stop-gradient, such that $\nabla_\theta \mathcal{L}_{\mathcal{M}} = 0$ for the visual encoder.

This mechanism addresses two interconnected challenges:

1) **Trade-off Resolution** Without gradient control, localization gradients (favoring high-frequency details) and detection gradients (favoring semantic abstraction) interfere within shared parameters, creating performance trade-offs. Controlled gradient flow enables the encoder to learn representations supporting both semantic and frequency-sensitive objectives simultaneously.
2) **Generalization and Robustness** In complex, multi-level manipulations containing diverse artifacts, unified optimization often amplifies gradient interference. By protecting the semantic learning path from excessive high-frequency updates, our approach maintains stable training dynamics and ensures consistent performance across varied manipulation types.

Through systematic experimentation (Section V-C), we determine $\lambda = 0.8$ as the optimal scaling factor that balances localization learning with semantic preservation.

E. TASK-SPECIFIC HEADS

Following dual-branch feature extraction, we employ three specialized heads that leverage task-appropriate features determined by the gradient regulation via scaling. Each head projects task-specific embeddings to prediction space through dedicated multi-layer perceptrons (MLPs).

1) BINARY CLASSIFICATION HEAD

The detection head processes spatial features $\mathbf{E}_{\text{spatial}}$ to preserve semantic coherence. Cross-modal fusion occurs through multi-head cross-attention between visual and textual representations:

$$\mathbf{h}_{\text{cls}} = \text{CrossAttn}(\mathbf{E}_{\text{spatial}}, \mathbf{E}_{\text{text}})_{[\text{CLS}]} \in \mathbb{R}^D \tag{5}$$

where the output embedding at the classification token position [CLS] aggregates cross-modal evidence into a D-dimensional vector. This fused representation is projected to binary logits through a two-layer MLP:

$$\hat{y}_{\mathcal{D}} = \text{MLP}_{\text{cls}}(\mathbf{h}_{\text{cls}}) \in \mathbb{R}^2 \tag{6}$$

The classification loss employs cross-entropy over the predicted distribution:

$$\mathcal{L}_{\mathcal{D}} = \text{CE}(\text{softmax}(\hat{y}_{\mathcal{D}}), y_{\mathcal{D}}) \tag{7}$$

where $y_{\mathcal{D}} \in \{0, 1\}$ denotes the ground-truth binary classification.

2) IMAGE GROUNDING HEAD

Bounding box regression leverages frequency-domain features fused with textual context through a learnable localization query:

$$\mathbf{h}_{\text{loc}} = \text{CrossAttn}(\mathbf{E}_{\text{freq}}, \mathbf{E}_{\text{text}})_{[\text{LOC}]} \in \mathbb{R}^D \tag{8}$$

where the localization query token [LOC] attends to both frequency-enhanced visual features and textual context. The resulting D-dimensional embedding is decoded into normalized box coordinates through a three-layer MLP with ReLU activations:

$$\mathbf{b} = \text{sigmoid}(\text{MLP}_{\text{bbox}}(\mathbf{h}_{\text{loc}})) \in [0, 1]^4 \tag{9}$$

where $\mathbf{b} = (x, y, w, h)$ represents the predicted box with normalized center coordinates (x, y) and dimensions (w, h) relative to image size. The sigmoid activation ensures all values lie in $[0, 1]$. The localization loss combines L1 regression and GIoU supervision:

$$\mathcal{L}_{\mathcal{B}} = \lambda_{\text{L1}} \|\mathbf{b} - \hat{\mathbf{b}}\|_1 + \lambda_{\text{GIoU}} \mathcal{L}_{\text{GIoU}}(\mathbf{b}, \hat{\mathbf{b}}) \tag{10}$$

where $\hat{\mathbf{b}}$ denotes ground-truth boxes. The L1 term provides coordinate-level gradients crucial for non-overlapping predictions, while GIoU optimizes spatial alignment through:

$$\mathcal{L}_{\text{GIoU}} = 1 - \left(\text{IoU}(\mathbf{b}, \hat{\mathbf{b}}) - \frac{|C \setminus (B \cup \hat{B})|}{|C|} \right) \tag{11}$$

where C denotes the smallest enclosing box containing both predicted B and ground-truth $\hat{B}$ boxes.

3) TEXT GROUNDING HEAD

Token-level manipulation detection uses frequency-domain features as visual context with stop-gradient to isolate text encoder optimization. The text encoder performs cross-attention with stopped gradients from visual features:

$$\mathbf{H}_{\text{token}} = \text{TextEncoder}(\text{sg}(\mathbf{E}_{\text{freq}}), \mathbf{E}_{\text{text}}) \in \mathbb{R}^{L\times D} \tag{12}$$

where $\text{sg}(\cdot)$ prevents gradients from $\mathcal{L}_{\mathcal{M}}$ affecting visual encoder parameters (Eq. 4), and $\mathbf{H}_{\text{token}}$ contains L token embeddings. Each token embedding is independently classified through a shared MLP:

$$\hat{y}_j = \text{MLP}_{\text{token}}(\mathbf{h}^j_{\text{token}}) \in \mathbb{R}^2, \quad j = 1, \ldots, L \tag{13}$$

The token grounding loss averages cross-entropy over all tokens:

$$\mathcal{L}_{\mathcal{M}} = \frac{1}{L}\sum_{j=1}^{L} \text{CE}(\text{softmax}(\hat{y}_j), y_j) \tag{14}$$

where $y_j \in \{0, 1\}$ indicates whether token j is manipulated. This architectural design ensures that token-level supervision focuses on text encoder parameters while utilizing frequency-domain visual context.

a: JOINT OPTIMIZATION

The total training objective combines all task-specific losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{D}} + \mathcal{L}_{\mathcal{B}} + \mathcal{L}_{\mathcal{M}} \tag{15}$$

allowing the gradient regulation via scaling (Eq. 4) to naturally balance optimization without manual loss weighting.

V. EXPERIMENTS

A. EXPERIMENTAL SETUP

Datasets We evaluate on: (1) **DGM4** [9]: About 23k image-text pairs with single-level manipulations (FS, FA, TS, TA); (2) **DGM4-Complex**: About 27k image-text pair extension with multi-level manipulations (29.29%). For training, we remove duplicated `orig` entries that arise from pairwise construction with multiple manipulation variants. This preprocessing step affects only the training split and does not modify the benchmark statistics or evaluation splits.

After this preprocessing, the training, validation, and test splits of DGM4-Complex contain 189,339, 26,519, and 61,073 image-text pairs, respectively. The validation and test splits serve as the evaluation sets and are not affected by training-time deduplication. The total of 276,931 pairs across all splits matches the full benchmark statistics reported in Table 4.

Implementation Built on HAMMER++ [10]: Vision Transformer (ViT)-B/16 visual encoder (256 × 256), BERT-base text encoder (50 tokens), single-level Haar DWT with 1 × 1 fusion. Training: AdamW optimizer, learning rates 3×10^{-5} (backbone) and 1×10^{-3} (heads), gradient scaling $\lambda = 0.8$, 50 epochs with cosine scheduling, batch size 32. Hardware: 4× NVIDIA A100 GPUs.

Metrics Binary Classification: AUC, EER, ACC; Image Grounding: IoU_m, IoU@50, IoU@75; Text Grounding: Token-level Precision, Recall, F1. Relative degradation is computed as (Complex − DGM4)/DGM4 × 100%.

B. MAIN RESULTS

Table 7 presents comprehensive performance comparison across manipulation complexity levels. On single-level manipulations (DGM4), CSCL achieves 96.34% AUC compared to GraDual 93.28%, indicating stronger performance under single-level manipulation settings. UFAFormer [12], which addresses the trade-off through frequency-aware architectural modules, reports 93.81% AUC on DGM4; our approach differs by targeting the optimization dynamics via differentiated gradient regulation across task heads. Performance patterns shift notably on multi-level manipulations (DGM4-Complex), with GraDual achieving 84.78% AUC against CSCL's 79.23%, reflecting enhanced robustness to compound forgery patterns.

Analysis of degradation patterns reveals distinct robustness characteristics. GraDual exhibits 9.1% performance reduction when transitioning from single to multi-level scenarios, compared to 17.8% for CSCL and 15.9-17.4% for other baselines. Text grounding tasks demonstrate notable stability, with GraDual maintaining 0.4% degradation while baseline methods range from 6.4% to 12.1%. Table 7 reports comprehensive metrics on DGM4-Complex: 84.78% AUC for detection, 68.94% IoU_m for image grounding, and 71.74% F1 for text grounding. These results suggest that gradient-controlled optimization contributes to improved robustness across varying manipulation complexities compared to unified encoder architectures.

Per-category analysis Table 8 reports per-category performance on the DGM4-Complex test split. GraDual shows the largest gap on TAM (82.75% vs 69.10% for HAMMER++ and 67.20% for CSCL, over 13pp) and on NT_{CTRL} false-positive rate (10.2% vs 20.8% and 20.5%). On NT, GraDual (87.27%) is close to CSCL (86.10%) and ahead of HAMMER++ (80.80%). On FAM and single manipulation categories, all methods perform within a few percentage points of each other. The pattern is consistent with the design of GraDual, which targets text-related compound manipulations and false-positive robustness through stop-gradient and gradient regulation on the shared encoder.

Computational complexity Table 9 compares computational cost. Training takes ∼48h for 50 epochs on 4×A100, longer than HAMMER++ (∼31h) and CSCL (∼23h) due to dual-pass feature extraction. Inference, however, stays close to HAMMER++ at 2.17ms per sample (vs 1.95ms), and runs much faster than CSCL (11.01ms). FLOPs (59.00G) are also comparable to HAMMER++ (54.12G).

External dataset evaluation To verify that the multi-level robustness of GraDual reflects its design rather than benchmark bias, we evaluate all three methods on three independent

TABLE 7. Performance comparison across manipulation complexity levels. Relative degradation (Δ) is computed as the percentage change from DGM4 to DGM4-Complex. Best results per dataset are shown in bold, and the smallest degradation is underlined. †Evaluated on DGM4 only due to unavailable code and checkpoints.

Method	Binary Classification (AUC %)			Image Grounding (IoUmean %)			Text Grounding (F1 %)		
	DGM4	DGM4-Complex	Δ (rel., %)	DGM4	DGM4-Complex	Δ (rel., %)	DGM4	DGM4-Complex	Δ (rel., %)
CLIP [62]	83.22	68.75	*-17.4*	49.51	36.82	*-25.6*	32.03	24.22	*-24.4*
ViLT [63]	85.16	71.26	*-16.3*	59.32	48.15	*-18.8*	57.00	49.65	*-12.9*
HAMMER++ [10]	93.19	78.41	*-15.9*	76.45	63.12	*-17.4*	71.35	66.81	*-6.4*
UFAFormer† [12]	93.81	–	–	78.33	–	–	72.02	–	–
CSCL [11]	**96.34**	79.23	*-17.8*	**84.07**	64.45	*-23.3*	76.62	67.38	*-12.1*
GraDual (Ours)	93.28	**84.78**	**-9.1**	76.88	**68.94**	**-10.3**	72.00	**71.74**	**-0.4**

TABLE 8. Per-category performance on DGM4-Complex test split. Values are AUC (%) unless noted. NT_{CTRL} reports FPR (%, lower is better). Best results per category in bold.

Category	N	GraDual	HAMMER++	CSCL
Single manipulation	37,862	92.05	91.88	**94.75**
FAM	7,942	85.12	**86.20**	84.90
TAM	7,329	**82.75**	69.10	67.20
NT	4,886	**87.27**	80.80	86.10
NT_{CTRL} (FPR↓)	3,054	**10.2**	20.8	20.5
Overall	61,073	**84.78**	78.41	79.23

TABLE 9. Computational complexity comparison. Inference with batch size 32, training over 50 epochs on 4×A100.

			Inference		Training (50 ep)	
Method	Params (M)	FLOPs (G)	Time (ms)	Mem (MB)	Total time (h)	Mem (MB)
HAMMER++	442.67	54.12	1.95	2,298	∼31	14,268
CSCL	461.92	77.46	11.01	2,316	∼23	25,370
GraDual (Ours)	454.48	59.00	2.17	2,448	∼48	20,894

TABLE 10. Lambda sensitivity sweep on the DGM4-Complex validation split. λ is varied from 0 to 1 in steps of 0.1. The best configuration per metric is shown in bold.

λ	AUC (%)	IoU (%)	F1-tok (%)
0.0	81.92	69.01	47.77
0.1	81.60	69.18	48.14
0.2	81.27	69.34	48.51
0.3	81.03	69.57	48.20
0.4	80.79	69.79	47.88
0.5	81.52	69.64	48.69
0.6	81.90	68.21	47.42
0.7	84.23	67.72	48.69
0.8	**84.46**	68.20	**50.93**
0.9	82.48	69.11	49.83
1.0	82.06	**70.02**	48.72

benchmarks (FaceForensics++, DFDCP, CelebDF) under short fine-tuning. All three methods reach reasonable performance (above 90% AUC across all settings), confirming that the performance gaps observed on DGM4-Complex are not due to benchmark construction. Full results are provided in the Appendix (Table 14).

TABLE 11. Ablation on text head gradient suppression (λ_{text}) on the DGM4-Complex validation split, with $\lambda = 0.8$ fixed. Stop-gradient ($\lambda_{text} = 0$) achieves the best performance on AUC and F1-tok. The best result per metric is shown in bold.

λ_{text}	AUC (%)	IoU (%)	F1-tok (%)
0.0 (stop-grad)	**84.46**	68.20	**50.93**
0.05	81.07	**69.83**	46.19
0.1	81.83	68.08	47.66
0.2	81.48	68.07	48.20

C. ABLATION STUDY

Lambda sensitivity

We conduct a comprehensive sweep over λ from 0 to 1 in steps of 0.1 (eleven values) on the DGM4-Complex validation split, reported in Table 10. Two boundary cases provide context. With $\lambda = 1.0$, localization gradients fully propagate to the shared encoder, eliminating the gradient regulation component of GraDual. With $\lambda = 0.0$, localization gradients are blocked, removing artifact-sensitive supervision from the encoder. The design intent of GraDual is to preserve substantial localization gradient flow while reducing interference with semantic representations, reflected in the choice of $\lambda = 0.8$ which retains 80% of the original gradient magnitude. The sweep confirms this choice empirically. $\lambda = 0.8$ achieves the best result on AUC (84.46%) and F1-tok (50.93%), and is therefore adopted as the operating point. AUC exhibits a sharp transition near $\lambda = 0.7$–0.8, suggesting that substantial gradient flow from the localization head ($\geq$ 70%) is required to engage frequency-sensitive representations meaningfully within the shared encoder; below this threshold the encoder behaves closer to the spatial-only configuration, while above $\lambda = 0.8$ the gradient regulation effect attenuates as expected. The corresponding test set performance is reported in Table 7. The slight numerical difference between validation (Table 10) and test (Table 12) performance at $\lambda = 0.8$ reflects natural variation across splits. The optimal λ selection is consistent between the two.

Stop-gradient on text head To verify the stop-gradient design choice ($\lambda_{text} = 0$) for the text grounding head, we compare against partial suppression with

TABLE 12. Ablation study on DGM4-Complex. Relative improvements (Δ) are measured against the Baseline. Gradient control resolves the semantic–frequency trade-off, yielding consistent gains across tasks.

Method	Performance			Δ (to baseline)		
	AUC (%)	IoUmean (%)	F1-tok (%)	AUC	IoU	F1-tok
Baseline	78.41	63.12	50.86	–	–	–
+ frequency features	80.53	68.96	47.70	+2.70%↑	+9.24%↑	−6.21%↓
+ Dual-path (λ=0.5)	83.21	68.35	53.05	+6.12%↑	+8.29%↑	+4.30%↑
+ Dual-path (λ=0.8)	**84.78**	70.94	53.48	+8.12%↑	+12.39%↑	+5.15%↑
+ Dual-path (λ=1.0)	81.73	70.45	51.61	+4.23%↑	+11.61%↑	+1.47%↑

TABLE 13. AUC (%) on DGM4 single-level test for the dual-branch and stop-gradient ablation, with the visual encoder frozen and the paper configuration in bold.

Text-head gradient scale	w/o DWT	w/ DWT
0 (stop-gradient)	92.63	**93.28**
0.05	92.63	93.17

TABLE 14. External dataset evaluation. Frame-level AUC (%). †HAMMER++ on FF++ uses encoder freezing due to fine-tuning instability with its default learning rate.

Method	FF++	DFDCP	CelebDF
HAMMER++ [10]	93.63†	95.79	90.45
CSCL [11]	98.55	95.32	93.99
GraDual (Ours)	97.96	92.83	92.28

$\lambda_{\text{text}} \in \{0.05, 0.1, 0.2\}$, reported in Table 11. Stop-gradient achieves the best performance on AUC (84.46%) and F1-tok (50.93%). Partial suppression degrades F1-tok by 2.73–4.74pp and AUC by 2.63–3.39pp across all tested configurations, indicating that allowing gradient flow from the text head impairs both text grounding and detection. Although $\lambda_{\text{text}} = 0.05$ yields a marginal IoU improvement of 1.63pp, this metric is governed by the separate $\lambda = 0.8$ scaling on the image grounding head, and stop-gradient remains the preferred design as it maximizes the two primary metrics (AUC and F1-tok) tied to the text grounding objective. The result empirically validates stop-gradient as the design choice for isolating text grounding optimization from the shared visual encoder.

Table 12 validates the effectiveness of **GraDual**, which integrates a dual-branch architecture with gradient scaling on DGM4-Complex. Frequency features alone introduce the trade-off identified in Section IV-B: IoU improves (+9.24%) while F1-token degrades (−6.21%). The dual-branch architecture with $\lambda = 0.8$ resolves this conflict, achieving simultaneous improvements of +8.12% AUC, +12.39% IoU, and +5.15% F1-token over the baseline. To isolate the contribution of gradient scaling, we further evaluate the dual-branch architecture without gradient regulation ($\lambda = 1.0$). This configuration yields a +4.23% AUC gain over the baseline, substantially smaller than the +8.12% achieved with $\lambda = 0.8$, indicating that differentiated gradient regulation is a key contributor to the detection performance gain. All reported improvements are relative percentages computed with respect to the baseline model.

D. ANALYSIS

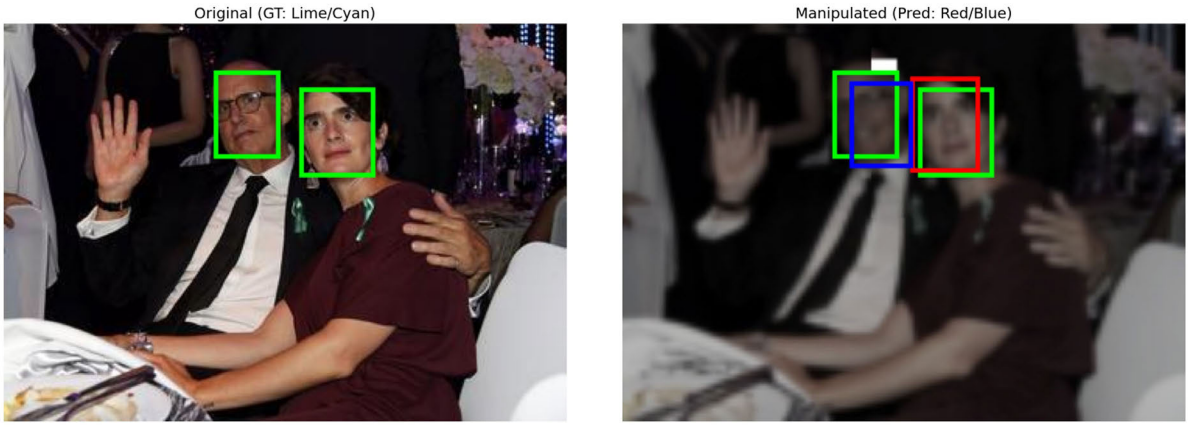
Performance Characteristics GraDual specializes in complex manipulation scenarios. While CSCL achieves 96.34% AUC on single-level DGM4 versus GraDual 93.28%, this reverses on DGM4-Complex (84.78% vs. 79.23%). Text grounding stability (72.00%→71.74%, −0.4%) particularly demonstrates the effectiveness of stop-gradient isolation.

Single-level performance gap

On DGM4, GraDual achieves 93.28% AUC, on par with HAMMER++ (93.19%) and UFAFormer (93.81%), while CSCL reports 96.34%. The dual-branch design is not the source of the gap. Both single-branch (HAMMER++) and dual-branch (UFAFormer, GraDual) methods reach comparable single-level performance, with frequency-aware models even slightly higher. CSCL's higher single-level AUC reflects its specialization through contrastive learning, which yields strong performance on simple manipulations but degrades sharply under compound forgery patterns (17.8% on DGM4-Complex versus GraDual's 9.1%). A direct ablation of the dual-branch and stop-gradient design choices on DGM4 single-level shows variations within 0.65 pp of the paper configuration (Table 13), in line with this observation.

Trade-off Resolution Gradient control with $\lambda = 0.8$ scales localization gradients to 80% of their original magnitude, allowing artifact-sensitive updates while reducing interference with semantic representations. This optimization-level approach differs from architectural fusion methods by directly addressing gradient conflicts.

Generalization and Robustness GraDual maintains stable optimization through dedicated pathways, exhibiting consistently smaller performance drops across complexity



Original Text:

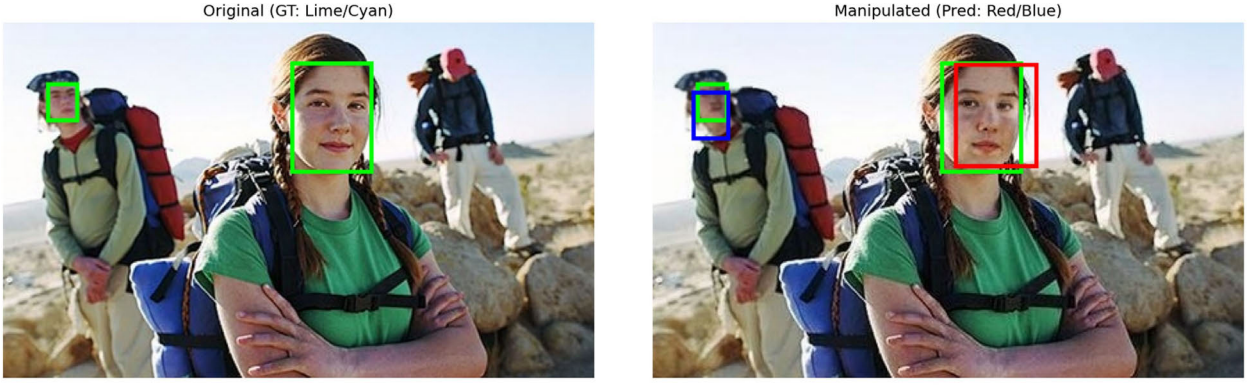
Actors Jeffrey Tambor and Gaby Hoffmann attend the Emmy Awards Governors Ball in Los Angeles

Manipulated Text Analysis:

Actors Jeffrey Tambor and Gaby Hoffmann attend the Emmy Awards Governors Ball in Los Angeles, including four people

Original Text:

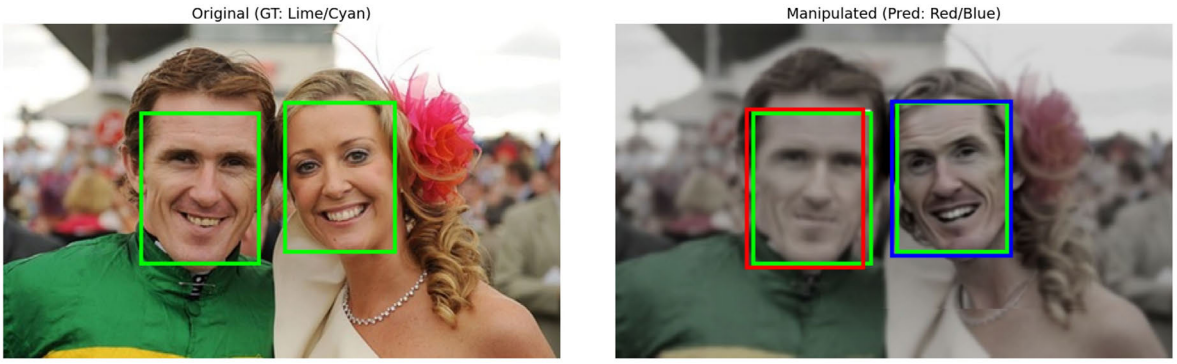
Newly elected Ukip MPs Mark Reckless and Douglas Carswell Ukip has fed on hatred of workings of Westminster



Original Text:

James Derounian considers a moneysaving green gap year for students

FIGURE 4. **Inference examples. From top to bottom: model predictions for samples 1–3.**

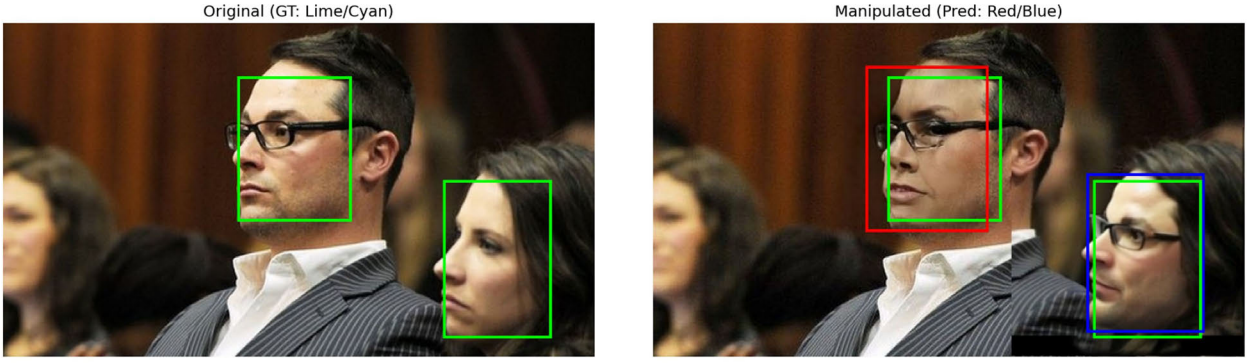


Original Text:

Tony McCoy and his wife Chanelle were left amused by a taxi driver who asked if they had heard the day s big racing news

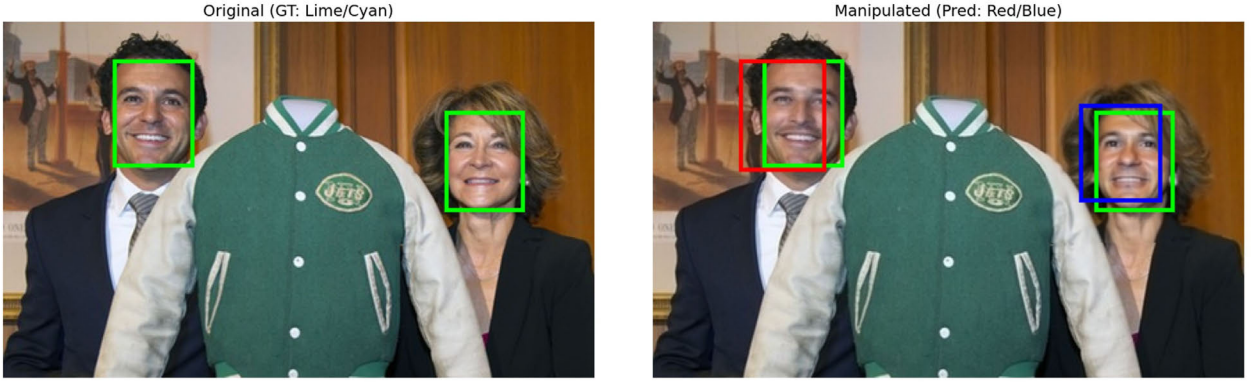
Manipulated Text Analysis:

Books on Mark Webber Steven Gerrard Michael Lynagh and Tony McCoy are featured among the best reads of the year



Original Text:

Carl Pistorius alongside his sister Aimee has been a nearconstant presence at his brother s murder trial



Original Text:

Fred Savage and his mother Joanne Savage pose with his character s Jets jacket

Manipulated Text Analysis:

Sentimental but satisfying Luke Edwards and Fred Savage in The Wizard

FIGURE 5. **Inference examples. From top to bottom: model predictions for samples 4–6.**

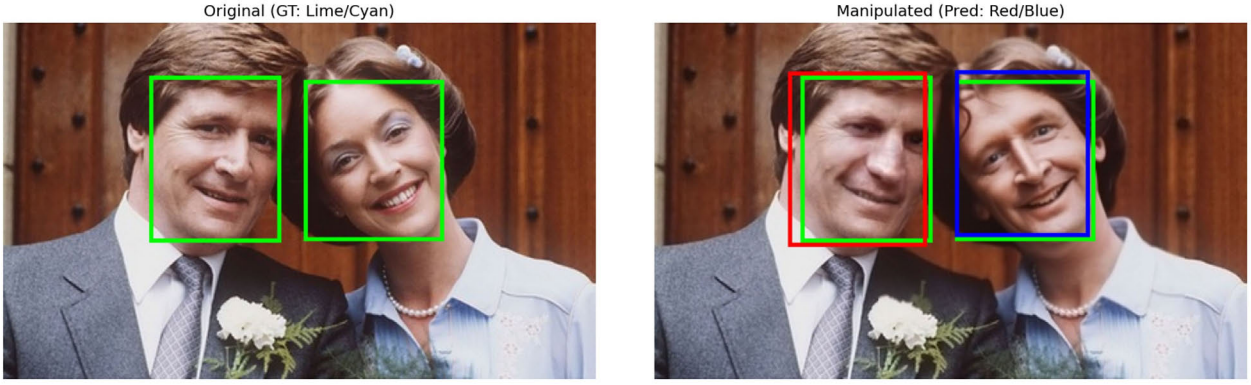
levels. Methods with pronounced trade-offs in ablation studies demonstrate larger degradation, confirming that unresolved gradient interference amplifies under manipulation diversity.

Original Text:

Georgia s ruling coalition s presidential candidate Georgy Margvelashvili with his daughter Anna outside a polling station in Tbilisi

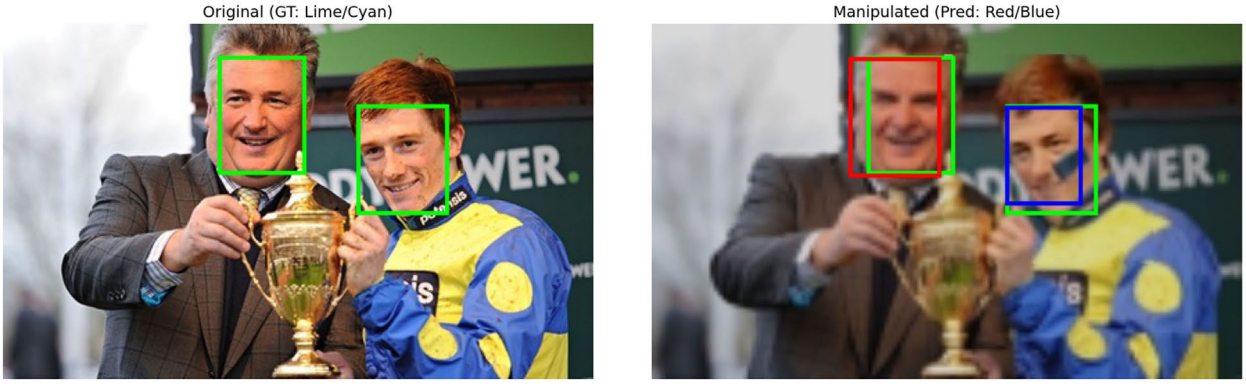
Manipulated Text Analysis:

People attend a preelection concert to support Georgy Margvelashvili, with four people in attendance



Original Text:

Ken Barlow and Deirdre Langton tie the knot in 1981



Original Text:

Paul Nicholls left and Sam TwistonDavies hold the Paddy Power Gold Cup following the victory of Caid Du Berlais

FIGURE 6. **Inference examples. From top to bottom: model predictions for samples 7–9.**

VI. CONCLUSION AND DISCUSSIONS

This paper addresses the challenge of detecting multi-level manipulations in multimodal content, where existing methods struggle to generalize effectively. We introduce DGM4-Complex, a comprehensive benchmark that incorporates auxiliary manipulations to simulate realistic multi-level

manipulation cases, overcoming the key limitation of existing datasets that contain only single manipulations per modality. To address this conflict, we propose GraDual (Gradient-Regulated Dual-branch), a gradient-controlled framework that balances semantic and frequency learning through differentiated gradient propagation, enabling stable multi-task optimization.

Experimental results show that GraDual performs consistently under complex manipulation scenarios. GraDual achieves an AUC of 84.78% in DGM4-Complex with only a 9.1% decrease from single-level cases, substantially lower than the 15.9–17.8% drop observed in existing methods. The text grounding task further highlights this robustness, as GraDual maintains nearly stable performance (72.00%→ 71.74%, a 0.4% decrease), whereas prior methods experience notable degradation of 6.4–12.9%. This stability suggests that isolating textual optimization from visual feature updates contributes to improved robustness. Furthermore, the ablation results indicate that gradient control mitigates the semantic–frequency trade-off, enabling simultaneous relative improvements of +8.12% AUC, +12.39% IoU, and +5.15% F1-token over the baseline. These findings highlight the importance of optimization-aware design for multimodal manipulation detection systems.

A. LIMITATIONS AND FUTURE WORK

While our method achieves strong performance on multi-level manipulations, the slightly lower performance on single-level cases suggests room for further optimization. In addition, the framework has not yet been evaluated on unseen or more diverse manipulation techniques beyond those included in DGM4-Complex. Future work includes extending the framework to video and audio modalities where temporal dynamics introduce additional challenges, evaluating robustness against emerging manipulation methods, and exploring adaptive gradient scaling in which λ dynamically adjusts according to manipulation complexity. Future work may explore adaptive architectures that first estimate the presence of frequency-domain artifacts and dynamically switch between single-branch and dual-branch processing.

APPENDIX
DGM4-COMPLEX IMPLEMENTATION DETAILS
A. FACE AUXILIARY MANIPULATION (FAM)

FAM synthesizes locally manipulated faces by conditioning a diffusion model on visual and semantic priors while preserving the global context of the target image. The pipeline consists of five computational stages.

- **Face Detection and Alignment**
  Faces are detected and cropped into 512×512 windows. A similarity transform $\mathbf{A}$ derived from five landmarks (eye corners, nose tip, mouth corners) aligns both source and target faces:

$$\mathbf{I}_{256} = \mathcal{W}(\mathbf{I}_{orig}, \mathbf{A})$$

- **Feature Extraction**
  Two feature embeddings are extracted
  - identity embedding $\mathbf{z}_{id}$
  - visual embedding $\mathbf{z}_{vis}$
- **Condition Preparation**
  Landmarks are converted to Gaussian heatmaps

$$C_{pts}(x, y) = \sum_i \exp\left(-\frac{\|(x, y) - p_i\|^2}{2\sigma^2}\right)$$

  The ControlNet conditioning tensor

$$I_{cond} = [C_{pts}; M_{face}]$$

- **Diffusion-Based Face Generation**
  ControlNet injects conditioning signals

$$h^l_{unet} \leftarrow h^l_{unet} + \alpha_t f^l_{ctrl}(I_{cond}, t)$$

  Classifier-free guidance

$$\epsilon_\theta = \epsilon^{(u)}_\theta + s(\epsilon^{(c)}_\theta - \epsilon^{(u)}_\theta)$$

- **Compositing**

$$I_{out} = M_{face} \odot \hat{I}_{face} + (1 - M_{face}) \odot I_{target}$$

B. TEXT AUGMENTATION

Three deterministic strategies introduce multimodal inconsistencies.

1) TAM: TEXT-AUDIENCE MISMATCH

$$n_{fake} = \max(n_{true} + \delta, 3), \quad \delta \in \{1, 2, 3\}$$

Phrase templates:

- with N people in attendance
- including N people
- among N attendees

2) NT: NONSPECIFIC TAIL

Candidate phrases:

- with updates pending
- with limited context
- with no further information available
- with an official summary unavailable
- with background details undisclosed

3) NT_CTRL

Applied only to samples where:

`fake_cls = orig`

This allows measurement of stylistic bias where models might incorrectly associate surface text patterns with manipulation signals.

APPENDIX
GRADUAL IMPLEMENTATION DETAILS
A. DUAL-BRANCH VISUAL FEATURE EXTRACTION

Two visual pathways feed a shared ViT-Base encoder.

Spatial branch:

$$E_{spatial} = Encoder_\theta(I)$$

Frequency branch:

$$E_{freq} = Encoder_\theta(DWT(I))$$

B. GRADIENT REGULATION

Scaled gradient operator:

$$G(x, \lambda) = x, \qquad \frac{\partial G}{\partial x} = \lambda I$$

Stop-gradient operator:

$$sg(x) = x, \qquad \frac{\partial sg}{\partial x} = 0$$

Encoder update:

$$\nabla_\theta L_{total} = \nabla_\theta L_{det} + \lambda \nabla_\theta L_{loc}$$

APPENDIX
EXTERNAL DATASET EVALUATION

To assess transferability and rule out benchmark construction bias, we fine-tune GraDual, HAMMER++, and CSCL on three independent benchmarks (FaceForensics++, DFDCP, CelebDF) under a short training schedule (3-5 epochs). Each method uses its native protocol with the visual encoder learning rate from its original implementation. HAMMER++ on FF++ uses encoder freezing due to fine-tuning instability with its default high encoder learning rate.

Results in Table 14 show that all three methods reach above 90% AUC across the external benchmarks. This indicates that the substantial performance gaps observed on DGM4-Complex reflect GraDual's design for multi-level manipulation patterns rather than benchmark construction bias.

APPENDIX
ADDITIONAL INFERENCE EXAMPLES

The green bounding box in the left image indicates the ground-truth manipulated region. Blue and red boxes represent the model predictions. When the accompanying text contains a manipulation, manipulated tokens are highlighted in green while model predictions are shown with red underlines.

REFERENCES

[1] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.

[2] Karthika. S and M. Durgadevi, "Generative adversarial network (GAN): A general review on different variants of GAN and applications," in *Proc. 6th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jul. 2021, pp. 1–8.

[3] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, pp. 1–72, 2025.

[4] B. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Rev.*, vol. 107, pp. 1753–1819, Dec. 2019.

[5] J. Twomey, D. Ching, M. P. Aylett, M. Quayle, C. Linehan, and G. Murphy, "Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine," *PLoS ONE*, vol. 18, no. 10, Oct. 2023, Art. no. e0291668.

[6] M. Westerlund, "The emergence of deepfake technology: A review," *Technol. Innov. Manage. Rev.*, vol. 9, no. 11, pp. 39–52, Jan. 2019.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[8] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8692–8701.

[9] R. Shao, T. Wu, and Z. Liu, "Detecting and grounding multi-modal media manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6904–6913.

[10] R. Shao, T. Wu, J. Wu, L. Nie, and Z. Liu, "Detecting and grounding multi-modal media manipulation and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5556–5574, Aug. 2024.

[11] Y. Li, Y. Yang, Z. Tan, H. Liu, W. Chen, X. Zhou, and Z. Lei, "Unleashing the potential of consistency learning for detecting and grounding multi-modal media manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 9243–9252.

[12] H. Liu, Z. Tan, Q. Chen, Y. Wei, Y. Zhao, and J. Wang, "Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation," *Int. J. Comput. Vis.*, vol. 133, no. 3, pp. 1392–1409, Mar. 2025.

[13] R. Shao, T. Wu, and Z. Liu, "Detecting and recovering sequential DeepFake manipulation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 712–728.

[14] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu, "Hierarchical fine-grained image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3155–3165.

[15] Z. Yu, J. Ni, Y. Lin, H. Deng, and B. Li, "DiffForensics: Leveraging diffusion prior to image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 12765–12774.

[16] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and encoder–decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, Jul. 2019.

[17] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9543–9552.

[18] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2022.

[19] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.

[20] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[21] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7887–7896.

[22] A. Sarkar, H. Mai, A. Mahapatra, S. Lazebnik, D. A. Forsyth, and A. Bhattad, "Shadows don't lie and lines can't bend! generative models don't know projective geometry.. for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 28140–28149.

[23] J. Ricker, D. Lukovnikov, and A. Fischer, "AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 9130–9140.

[24] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on gradients: Generalized artifacts representation for GAN-generated images detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12105–12114.

[25] Y.-Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 86–103.

[26] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16312–16321.

[27] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20606–20615.

[28] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5000–5009.

[29] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for diffusion-generated image detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22445–22455.

[30] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2023, pp. 3418–3432.

[31] K. Sun, S. Chen, T. Yao, H. Liu, X. Sun, S. Ding, and R. Ji, "DiffusionFake: Enhancing generalization in deepfake detection via guided stable diffusion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 101474–101497.

[32] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24480–24489.

[33] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware adaptive transformer for generalizable synthetic image detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 10770–10780.

[34] C. Koutlis, S. Papadopoulos, and I. Kompatsiaris, "RINE: Leveraging representations from intermediate encoder-blocks for synthetic image detection," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 467–484.

[35] R. Shao, T. Wu, L. Nie, and Z. Liu, "DeepFake-adapter: Dual-level adapter for DeepFake detection," *Int. J. Comput. Vis.*, vol. 133, no. 6, pp. 3613–3628, Jun. 2025.

[36] S. A. Khan and D.-T. Dang-Nguyen, "CLIPping the deception: Adapting vision-language models for universal deepfake detection," in *Proc. Int. Conf. Multimedia Retr.*, May 2024, pp. 1006–1015.

[37] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the bar of AI-generated image detection with CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 4356–4366.

[38] S. Aneja, C. Bregler, and M. Nießner, "COSMOS: Catching out-of-context image misuse using self-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 12, pp. 14084–14092.

[39] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14920–14929.

[40] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, "Sniffer: Multimodal large language model for explainable out-of-context misinformation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 13052–13062.

[41] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2823–2832.

[42] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other- audio-visual dissonance-based deepfake detection and localization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 439–447.

[43] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5037–5047.

[44] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10491–10503.

[45] T. Oorloff, S. Koppisetti, N. Bonettini, D. Solanki, B. Colman, Y. Yacoob, A. Shahriyari, and G. Bharaj, "AVFF: Audio-visual feature fusion for video deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27102–27112.

[46] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track*, 2021.

[47] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14918–14927.

[48] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18689–18698.

[49] S. Wu, J. Liu, J. Li, and Y. Wang, "Few-shot learner generalizes across AI-generated image detection," in *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, Jul. 2025, pp. 67449–67460.

[50] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22355–22366.

[51] Z. Chen, V. Badrinarayanan, C. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 794–803.

[52] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7482–7491.

[53] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 5824–5836.

[54] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, "Conflict-averse gradient descent for multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 18878–18890.

[55] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 527–538.

[56] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2003–2011.

[57] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3403–3412.

[58] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, "High-fidelity GAN inversion for image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11369–11378.

[59] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2065–2074.

[60] Y. Han, J. Zhu, K. He, X. Chen, Y. Ge, W. Li, X. Li, J. Zhang, C. Wang, and Y. Liu, "Face-adapter for pre-trained diffusion models with fine-grained id and attribute control," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2024, pp. 20–36.

[61] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[62] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[63] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

DONGHUN LEE received the B.S. degree in computer science from the University of Illinois Urbana–Champaign, USA, and the M.S. degree from Seoul National University, South Korea, where he is currently pursuing the Ph.D. degree. From 2012 to 2013, he was a Software Engineer with Google Korea. He is also a Researcher with the Artificial Intelligence Safety Institute, Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include multimodal learning, representation learning, video prediction, and latent world models.

JANG-HO CHOI received the B.S. degree in computer science and business administration from the University of Southern California, USA, in 2009, and the M.S. degree in computer science from KAIST, South Korea, in 2013. He is a Project Leader of a multimodal deepfake detection project and a Senior Researcher with the Artificial Intelligence Safety Institute, Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include AI safety, multimodal deepfake detection, explainable AI, and robust machine learning.

SEONGHO KIM (Associate Member, IEEE) received the B.S. degree in electronic engineering and the M.S. degree in electrical and computer engineering from Inha University, South Korea. He has been a Researcher with the Artificial Intelligence Safety Institute, Electronics and Telecommunications Research Institute (ETRI), South Korea, since 2025. His research interests include face editing, deepfake generation, and video manipulation.

SUNGWON YI received the M.S. and Ph.D. degrees in computer science and engineering from The Pennsylvania State University, in 2004 and 2005, respectively. Prior to his academic career, he was a Systems Engineer with LG-CNS, for four years. Since 2005, he has been with ETRI, Republic of Korea, where he is currently the Assistant Director of the AI Safety Research Institute. His research interests include machine learning, large language models, vision-language models, network security, and mobile computing. He has been an Active Member of the Technical Program Committee for IEEE GLOBECOM and ICC since 2005. He is the Editor-in-Chief of *ETRI Journal*.

NOJUN KWAK (Senior Member, IEEE) was born in Seoul, South Korea, in 1974. He received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, in 1997, 1999, and 2003, respectively. From 2003 to 2006, he was with Samsung Electronics, Seoul. In 2006, he joined Seoul National University, as a BK21 Assistant Professor. From 2007 to 2013, he was a Faculty Member with the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea. Since 2013, he has been with the Graduate School of Convergence Science and Technology, Seoul National University, where he is currently a Full Professor. His current research interests include feature learning by deep neural networks and their applications in various areas of pattern recognition, computer vision, image processing, and natural language processing.