UniGenDet: A Unified Generative-Discriminative Framework for Co-Evolutionary Image Generation and Generated Image Detection

Supplementary Material

A. Implementation Details

In this section, we provide comprehensive details regarding the training pipeline, model configurations, and inference settings to facilitate reproducibility.

A.1. Training Setup and Hyperparameters

The training process of UniGenDet is divided into two phases: Generation-Detection Unified Fine-tuning (GDUF) and Detector-Informed Generative Alignment (DIGA).

GDUF Stage. The complete training pipeline of the GDUF stage requires approximately 12 hours on 8 NVIDIA A100 GPUs. The model is trained for about 1,000 optimization steps with a total batch token count of 16384×8.

DIGA Stage. The subsequent DIGA training phase completes in approximately 6 hours over 500 steps s on 8 NVIDIA A100 GPUs. During this stage, we specifically align the features from the 8th layer of the generator with the final layer features of the detection module. The alignment loss weight is set to $\lambda = 0.5$. To ensure training stability and effectiveness, we maintain a balanced ratio of positive (real) and negative (fake) samples throughout the alignment process.

Data Preprocessing. Following the configuration of the base model BAGEL [2], we apply distinct preprocessing strategies for different tasks:

- **Detection Inputs:** Processed with a patch extraction strategy of [min=224, max=980, stride=14] using the SigLIP encoder.
- **Generation Inputs:** Processed with a patch extraction strategy of [min=512, max=1024, stride=16] for the VAE encoder.

A.2. Model Architectures

Our framework is built upon a powerful Large Language Model (LLM) and a high-resolution vision encoder. Detailed architectural configurations are listed in Table A. The visual encoder is based on a SigLIP [3] architecture, while the LLM backbone is initialized from Qwen2.5 [1] weights.

A.3. Inference Settings

For the text-to-image generation task, we employ a 50-step diffusion sampling process. For text generation (including detection explanations), we use nucleus sampling with a temperature of 0.7, top-p of 0.8, and top-k of 20. A repetition penalty of 1.05 is applied to prevent degenerative loops in the generated explanations.

Table A. **Detailed Model Configurations.** Specification of the Vision Encoder and LLM Backbone used in UniGenDet.

Component	Parameter	Value
LLM Backbone (Qwen2.5)	Hidden Size	3,584
	Intermediate Size	18,944
	Number of Layers	28
	Attention Heads	28
	Key/Value Heads	4
	Vocab Size	152,064
Vision Encoder (SigLIP)	Model Type	SigLIP Vision Model
	Image Size	980
	Patch Size	14
	Hidden Size	1,152
	Number of Layers	27

B. Attention Mask Mechanism

A core design of UniGenDet is the task-specific attention masking strategy that orchestrates the flow of information between text, visual understanding features, and generative latents.

Generation Task. In the text-to-image generation mode, the input sequence consists of text prompts followed by VAE latent noise.

- **Text Tokens:** We apply a *causal mask* (attending only to preceding tokens) to preserve the autoregressive nature of language modeling.
- **Visual Latents:** We employ a *bidirectional mask* within the image sequence, allowing image noise tokens to attend to all other image tokens and all preceding text tokens. This ensures the generated visual content aligns with the textual prompt.

Detection Task. The generated image detection task involves a more complex interaction among three components: the Generation Encoder (VAE), the Detection Encoder (ViT), and the Text Instructions.

- **VAE Latents (z_{gen}):** These tokens attend globally to themselves to model the generative distribution.
- **ViT Features (h_{det}):** Through the Symbiotic Multimodal Self-Attention (SMSA) module, these features attend globally to themselves and cross-attend to the VAE latents, enabling the detector to perceive generative artifacts.
- **Text Tokens:** The instruction and answer tokens use a *causal mask* internally. Importantly, they can attend to **all** preceding visual tokens (both VAE and ViT features) to perform grounded reasoning and generate explanations.

Figure A. **Comparison of generation result**. BAGEL (middle) vs UniGenDet (bottom) generation comparison. UniGenDet produces more natural landscapes with coherent lighting, validating detection-guided optimization.

Table B. Robustness Comparison under JPEG Compression.

JPEG Quality	Method	Accuracy↑	F1-Score↑	ROUGE-L↑	CSS↑
N/A	FakeVLM	**98.6**	**98.1**	32.2	59.5
	Ours	98.0	97.7	**56.3**	**79.8**
90	FakeVLM	86.8	86.4	29.5	54.2
	Ours	**95.1**	**94.7**	**54.0**	**76.7**
70	FakeVLM	88.4	88.0	30.2	54.8
	Ours	**91.4**	**90.8**	**52.1**	**75.1**
50	FakeVLM	80.4	80.3	29.3	53.2
	Ours	**91.3**	**90.7**	**51.7**	**74.4**

Table C. Robustness Comparison under Image Cropping.

Crop Ratio	Method	Accuracy↑	F1-Score↑	ROUGE-L↑	CSS↑
N/A	FakeVLM	**98.6**	**98.1**	32.2	59.5
	Ours	98.0	97.7	**56.3**	**79.8**
0.9	FakeVLM	95.4	95.0	30.6	57.5
	Ours	**97.7**	**97.5**	**55.2**	**78.4**
0.7	FakeVLM	93.8	93.3	30.9	57.8
	Ours	**97.3**	**97.1**	**54.6**	**78.0**
0.5	FakeVLM	92.3	91.8	30.0	57.7
	Ours	**95.4**	**95.0**	**52.6**	**77.0**

C. Robustness Analysis

We further evaluate the robustness of UniGenDet against common image perturbations, specifically JPEG compression and image cropping, which are frequent in social media dissemination. We compare our method with the state-of-the-art MLLM-based detector, FakeVLM.

JPEG Compression. As shown in Table B, UniGenDet exhibits superior stability. Even under severe compression (Quality=50), our method maintains an accuracy of 91.3%, surpassing FakeVLM by over 10%. This indicates that our model learns semantic-level forgery cues rather than relying solely on fragile high-frequency artifacts.

Image Cropping. Table C presents the performance under varying crop ratios. Our method demonstrates high resilience, maintaining 97.7% accuracy at a 0.9 crop ratio. This suggests that the unified training enables the model to identify local inconsistency effectively, even when global context is partially missing.

D. Qualitative Visualization

As depicted in Figure A, a comparative analysis of generation quality between BAGEL and our method is presented. The samples from BAGEL reveal noticeable artificiality, such as the disproportional pagoda against Mount Fuji and inconsistent lighting on the lake's surface. In contrast, the results from our proposed UniGenDet, refined via the authenticity-aware fine-tuning strategy, exhibit significantly enhanced realism. This improvement is evident in the coherent shadow transitions of the mountain, the natural water reflections, and the detailed texture of the character's attire, collectively demonstrating superior adherence to physical realism.

E. Differences from GANs and Generation Diversity

While UniGenDet utilizes a discriminative module to guide generation, its optimization fundamentally differs from Generative Adversarial Networks (GANs). GANs rely on a zero-sum game, which often leads to adversarial mode collapse. In contrast, our unified co-optimization paradigm

Table D. Generated Image Diversity. CLIP: CLIP Similarity, LPIPS: Learned Perceptual Image Patch Similarity

Method	UniGenDet	BAGEL
CLIP Similarity ↓ / LPIPS ↑	0.802 / 0.726	0.804 / 0.714

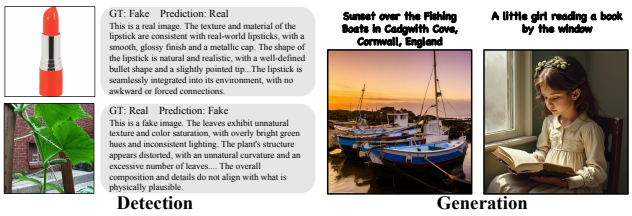

Figure B. Failure cases of detection and generation.

avoids this by utilizing constructive constraints. Instead of merely attempting to fool a classifier with binary scalar feedback, UniGenDet employs Detector-Informed Generative Alignment (DIGA) to perform explicit distribution alignment in a high-dimensional feature space. This approach stabilizes training by transferring continuous, rich forensic knowledge. Consequently, it prevents unnatural feedback loops, ensuring that even an imperfect detector constructively enhances physical plausibility without steering the generator into erroneous sub-spaces.

To empirically validate that UniGenDet does not suffer from mode collapse, we quantified generation diversity using 500 prompts from the LAION dataset, generating 16 variations for each prompt. As shown in Table D, UniGenDet achieves an average intra-prompt LPIPS of 0.726 and a CLIP similarity of 0.802. These metrics are highly comparable to the original BAGEL model, which yielded an LPIPS of 0.714 and a CLIP similarity of 0.804. These results confirm that our stabilizing feature-based optimization successfully preserves image diversity while improving overall synthesis quality.

F. Failure Cases Analysis

Despite its robust performance, UniGenDet encounters occasional limitations in both generation and detection, as visualized in Figure B. In the detection task, errors can occasionally occur when evaluating highly realistic forgeries or atypical real samples, such as heavily stylized or professionally post-processed authentic images. For the generation task, while the framework effectively corrects most physical anomalies, it may still yield inconsistent textures when synthesizing highly complex scenes. These failure cases suggest that future iterations of multi-modal foundational models could greatly benefit from incorporating more explicit fine-grained spatial reasoning and scaling up the diversity of the training distributions to handle extreme edge cases.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[2] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1

[3] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1