

Gated Residual Discrepancy for Efficient Bi-Encoder Factuality Verification

Zeyang Liu, Jingfeng Xue, Jiarun Fu, Wenbiao Du,
Xiuqi Yang, Yifeng Fu, Junbao Chen, Yong Wang*

Beijing Institute of Technology
{liuzeyang, wangyong}@bit.edu.cn

Abstract

Large Language Models (LLMs) often generate specious hallucinations, creating an urgent need for post-hoc factuality verification against reliable evidence. Existing verification methods incur substantial computational overhead due to joint evidence–claim encoding. While bi-encoders offer a highly efficient alternative, they struggle with asymmetric entailment and are severely vulnerable to contextual noise. To address these challenges, we propose Gated Residual Discrepancy (GRD), a highly efficient bi-encoder framework for factuality verification. To explicitly capture asymmetric entailment, GRD pioneers a difference-dominant paradigm, extracting dual-view vector differences. To mitigate the vulnerability to contextual noise, GRD employs an adaptive gating mechanism that dynamically filters contextual noise, preventing large semantic backgrounds from masking critical errors, thereby enabling highly sensitive verification over broad evidence contexts. Experiments across various benchmarks show GRD achieves 79.99% average Macro-F1, trailing the strongest cross-encoder baseline by only 1.15 points while delivering a $23.55\times$ speedup with only a fraction of the computational FLOPs. All datasets and code are publicly available.¹

1 Introduction

Large Language Models (LLMs) have achieved remarkable success in natural language generation, yet their propensity for factual hallucinations remains a critical barrier to their deployment in real-world applications that demand strict factual accuracy (Gekhman et al., 2024; Servadio et al., 2025; Fu et al., 2026; Yang et al., 2026). Such fluent but ungrounded outputs can propagate severe misinformation, mislead users, and severely undermine trust in AI systems, creating an urgent

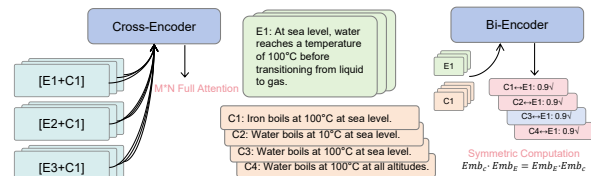


Figure 1: Cross-encoders suffer from expensive joint encoding. Traditional bi-encoders are efficient but rely on symmetric similarity, incorrectly assigning high scores to unsupported evidence–claim pairs due to substantial lexical overlap.

need to rigorously assess the truthfulness of generated content (Jiao et al., 2025; Du et al., 2026). To mitigate this issue, *post-hoc factuality verification* has emerged as a central research focus in natural language processing, aiming to verify whether generated claims are faithfully supported by evidence documents (Iqbal et al., 2024; Tang et al., 2024; Rahman et al., 2026).

Both LLM-based and cross-encoder verification methods fundamentally rely on joint encoding, requiring deep self-attention over concatenated evidence–claim pairs, as illustrated on the left of Figure 1 (Tang et al., 2024; Wei et al., 2024; Oh et al., 2025). Consequently, the lengthy evidence document must be redundantly encoded for every single claim. This inefficiency becomes computationally prohibitive in fine-grained verification, where generated text is decomposed into multiple atomic units to pinpoint errors. Ultimately, this compounded processing cost creates a severe throughput bottleneck, preventing large-scale deployment (Dorner et al., 2024).

A natural solution is to adopt bi-encoder architectures, which allow for offline evidence pre-computation and efficient embedding-space interaction (Devlin et al., 2019). However, as shown in Figure 1 (right), standard bi-encoders are highly susceptible to semantic interference due to their reliance on symmetric computation; ungrounded claims with high lexical overlap receive falsely

*Corresponding author.

¹<https://github.com/lzy248/grd>

high similarity scores, effectively masking critical factual contradictions. Furthermore, this symmetric paradigm fundamentally fails to capture the asymmetric entailment inherent in factuality verification, where support from a comprehensive evidence document to a specific claim does not imply the reverse direction. Recent methods like E-Verify (Liu et al., 2025) attempt to fuse symmetric concatenation features with asymmetric vector differences for factuality verification in embedding space, but these features are easily drowned out by contextual noise in lengthy evidence. To prevent subtle factual discrepancies from being drowned out, E-Verify adopts a dual-decomposition strategy that splits the evidence into atomic units. While this design improves sensitivity, it also introduces additional preprocessing overhead, weakening part of the efficiency advantage of the bi-encoder paradigm. Therefore, realizing an efficient bi-encoder framework requires overcoming two critical challenges: (1) *the inability to capture asymmetric entailment robustly*, and (2) *the severe vulnerability to contextual noise in long documents*.

To concurrently address these challenges, we propose Gated Residual Discrepancy (GRD), an efficient, noise-resilient, and asymmetric-aware bi-encoder framework for factuality verification. To address the first challenge, GRD adopts a difference-dominant discriminative paradigm. Moving away from the conventional reliance on feature fusion, GRD directly extracts residual signals via dual-view vector differences, capturing both direction and magnitude to represent asymmetric entailment. By relying solely on these vector differences, the framework avoids the semantic interference of symmetric features like concatenation, effectively preventing substantial topical overlap from masking subtle contradictions. To tackle the second challenge, GRD employs an adaptive gating mechanism to manage contextual noise directly in the embedding space. It dynamically distinguishes harmless lexical variations from severe factual clashes, selectively amplifying critical deviations while filtering out irrelevant linguistic noise. Consequently, it achieves robust verification against broad evidence contexts, eliminating the need for evidence-side atomic decomposition. Extensive evaluations across multiple benchmarks demonstrate that GRD achieves an average Macro-F1 of 79.99%, approaching the strongest cross-encoder baseline within 1.15 points, while deliver-

ing a $23.55\times$ speedup and reducing computational FLOPs by over two orders of magnitude.

Our contributions can be summarized as:

- We pioneer a difference-dominant discriminative paradigm through the GRD operator, extracting dual-view vector differences to robustly model asymmetric entailment.
- We introduce an adaptive gating mechanism that dynamically filters contextual noise in the embedding space. By preventing large semantic backgrounds from masking subtle factual errors, this internal denoising ensures highly sensitive verification.
- Extensive experiments show that GRD substantially reduces online computational FLOPs and accelerates inference by $23.55\times$ compared to joint-encoding methods, all while sustaining a cross-encoder-level average Macro-F1 of 79.99%.

2 Related Work

2.1 Factuality Verification Pipelines

Early factuality verification primarily evaluated generated content at the document or paragraph level. To provide precise error localization and enhanced interpretability, recent research has pivoted towards fine-grained verification, establishing the *decompose-then-verify* pipeline as the dominant paradigm (Min et al., 2023; Zhao et al., 2023; Stacey et al., 2024). In this framework, the generated text is first segmented into atomic claims, which are subsequently evaluated against the source evidence. Representative systems such as FActScore (Min et al., 2023), SAFE (Wei et al., 2024), and DiVA (Huang et al., 2026) all adopt this two-stage design. The verification stage typically relies on either Large Language Models (*LLM-as-judge*) or Natural Language Inference (*NLI-based*) cross-encoders (Laban et al., 2022; Pan et al., 2024). While LLM-based verifiers offer strong zero-shot capabilities, their large parameter sizes make inference costs prohibitive for large-scale deployment (Dorner et al., 2024). Consequently, NLI-based cross-encoders have emerged as a more practical alternative, significantly reducing computational overhead while maintaining highly competitive accuracy. However, both LLM and NLI-based verifiers fundamentally rely on pairwise joint encoding. In fine-grained scenarios, this architecture

forces the model to redundantly encode the lengthy evidence document for each atomic claim, creating a severe computational bottleneck that scales linearly with the number of decomposed units.

2.2 Embedding-based Entailment and Verification

Factuality verification is inherently an asymmetric entailment task. Historically, early representation learning for Natural Language Inference (NLI) attempted to capture this asymmetry by fusing independent sentence embeddings using heuristic concatenations, typically adopting forms like $[u; v; |u - v|]$ (Conneau et al., 2017; Reimers and Gurevych, 2019). This paradigm established feature fusion as the standard for embedding-based entailment. Recently, E-Verify (Liu et al., 2025) adapted this paradigm for factuality verification, pioneering the use of vector difference operations to capture logical contradictions. Despite these advances, existing embedding-based verifiers still face critical limitations. First, by continuing to fuse difference features with symmetric similarity metrics, they suffer from *similarity saturation*: in scenarios with high lexical overlap, the massive symmetric activation acts as background noise, masking subtle factual conflicts. Second, to mitigate this noise, current methods choose to decompose both evidence and claims into atomic units before encoding (Liu et al., 2025). This dual-decomposition not only inflates preprocessing overhead but also discards the holistic context of the source document, leaving the efficient and highly sensitive verification of unfragmented evidence an unresolved bottleneck in the bi-encoder paradigm.

3 Problem Formulation

Given a piece of evidence e and a claim c , the goal of factuality verification is to learn a discriminative function $f(e, c)$ that outputs a probability score in $[0, 1]$. This function estimates the probability that the factual content of c is faithfully supported by e , which naturally aligns with the asymmetric entailment objective in binary natural language inference (NLI).

Under the bi-encoder paradigm, the discriminative function f is decoupled into independent representation learning and embedding-space interaction. First, a shared text encoder $\mathcal{E}$ maps the discrete input texts into continuous embedding representations $\mathbf{H}_e = \mathcal{E}(e)$ and $\mathbf{H}_c = \mathcal{E}(c)$. Subse-

quently, a lightweight verification operator $\mathcal{V}$ computes the final entailment probability strictly within the embedding space. The entire pipeline is explicitly formulated as:

$$f(e, c) = \mathcal{V}(\mathbf{H}_e, \mathbf{H}_c) = \mathcal{V}(\mathcal{E}(e), \mathcal{E}(c)). \quad (1)$$

4 The GRD Framework

In this section, we introduce the Gated Residual Discrepancy (GRD) framework. While preserving the critical inference efficiency of bi-encoder architectures via offline pre-computation, GRD explicitly overcomes their inherent expressiveness bottlenecks. By concurrently addressing the inability to capture asymmetric entailment via a difference-dominant operator, and mitigating the severe vulnerability to contextual noise via an adaptive gating mechanism, GRD enables highly sensitive and robust verification.

4.1 Multi-Granularity Semantic Encoding

Conventional bi-encoder models compress an entire passage into a single [CLS] vector (Devlin et al., 2019). When dealing with lengthy, intact content, this creates a severe information bottleneck—fine-grained facts are easily overshadowed by dominant global semantics. To overcome this limitation and enable highly sensitive verification, we adopt a multi-view encoding strategy inspired by multi-vector representation learning (Lee et al., 2019; Khattab and Zaharia, 2020; Liao et al., 2024). By mapping each text into a fixed-length set of semantic slots, this approach explicitly expands the representational capacity, preserving diverse local facts for fine-grained discrepancy detection.

Specifically, given an input text x , we first extract token-level contextual representations $T_x = \{t_1, t_2, \dots, t_L\} \in \mathbb{R}^{L \times d}$ using a pre-trained language model (e.g., BERT), where L is the sequence length and d is the hidden dimension. We then apply a Pooling-based Multi-Head Attention (PMA) module to aggregate T_x into K slot vectors. PMA maintains K learnable seed vectors $S = \{s_1, \dots, s_K\} \in \mathbb{R}^{K \times d}$ as queries, and attends over the token sequence:

$$\hat{H} = \text{LN}(\text{MHA}(S, T_x, T_x) + S), \quad (2)$$

$$H_x = \text{LN}\left(\hat{H} + \text{FFN}(\hat{H})\right), \quad (3)$$

where MHA denotes Multi-Head Attention, LN denotes Layer Normalization, and FFN is a feed-forward network. The output $H_x \in \mathbb{R}^{K \times d}$ consists

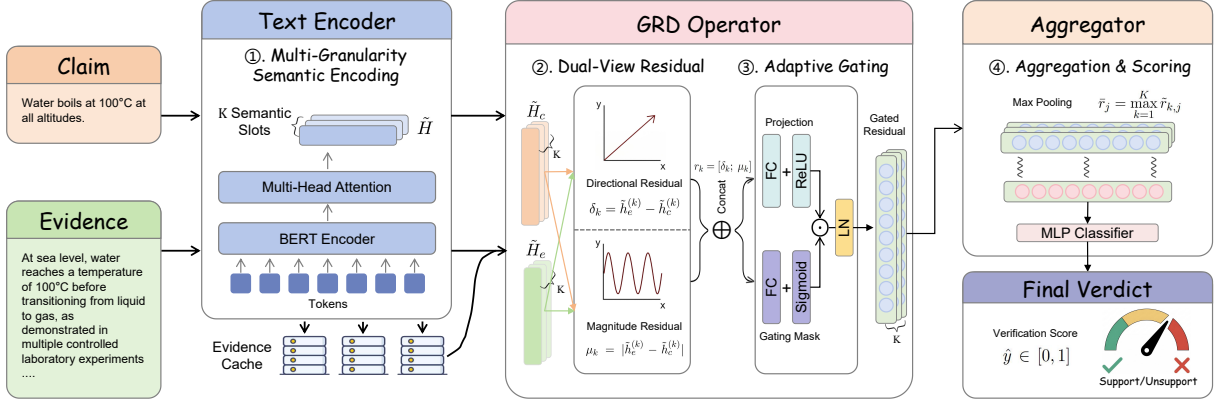


Figure 2: The overall architecture of the GRD framework. (1) The text encoder maps claims and evidence into K multi-granularity semantic slots, where evidence representations can be efficiently cached offline. (2) The GRD operator extracts dual-view residuals between corresponding slots to explicitly capture asymmetric entailment. (3) An adaptive gating mechanism filters contextual noise and amplifies critical factual deviations. (4) The aggregator applies max-pooling to preserve the strongest local contradiction signal for the final verification score.

of K vectors, denoted as $\{h_x^{(1)}, \dots, h_x^{(K)}\}$, each capturing a distinct semantic perspective of the input. Finally, for each $k \in \{1, \dots, K\}$, the slot vector is ℓ_2 -normalized to $\tilde{h}_x^{(k)} = h_x^{(k)} / \|h_x^{(k)}\|_2$.

By applying PMA independently to the evidence and the claim, we obtain their respective normalized multi-view representations $\tilde{H}_e = \{\tilde{h}_e^{(1)}, \dots, \tilde{h}_e^{(K)}\}$ and $\tilde{H}_c = \{\tilde{h}_c^{(1)}, \dots, \tilde{h}_c^{(K)}\}$. The K slots act as diverse semantic viewpoints, enabling the model to distribute heterogeneous information (e.g., temporal, entity, and action features) across separate vectors, thereby substantially raising representational resolution for local facts.

4.2 Dual-View Residual Extraction

Since the evidence and the claim are processed by the same multi-view encoder with shared PMA seeds, corresponding slots $\tilde{h}_e^{(k)}$ and $\tilde{h}_c^{(k)}$ are encouraged to capture comparable semantic aspects in a shared embedding space. This inherent alignment allows us to treat the direct vector difference between corresponding slots as a semantic residual that emphasizes factual discrepancies by attenuating shared content between the evidence and the claim.

Thus, for each slot $k \in \{1, \dots, K\}$, we extract two complementary residual views:

- **Directional residual:** $\delta_k = \tilde{h}_e^{(k)} - \tilde{h}_c^{(k)}$. This preserves the geometric direction of the shift in embedding space to explicitly model asymmetric entailment. Crucially, it captures not only in-place logical discrepancies (e.g., negation or entity substitution) but also the asymmetry of information coverage—encoding

whether the evidence possesses a semantic surplus that entails the claim, or an information deficit that exposes hallucinated content.

- **Magnitude residual:** $\mu_k = |\tilde{h}_e^{(k)} - \tilde{h}_c^{(k)}|$. While the directional view encodes the nature of the semantic shift, the magnitude view quantifies its severity. By measuring the element-wise absolute intensity of deviation, it provides a crucial gauge of discrepancy salience—enabling the model to differentiate between the low-magnitude fluctuations of benign paraphrasing and the glaring, high-magnitude spikes of critical factual contradictions.

The two views are concatenated to form a complete residual descriptor:

$$r_k = [\delta_k; \mu_k] \in \mathbb{R}^{2d}. \quad (4)$$

The directional and magnitude views are complementary: δ_k encodes *what* changed and *in which direction*, while μ_k encodes *how much* changed. Their combination provides a richer characterization of factual discrepancies than either view alone.

4.3 Adaptive Gating Amplification

When matching claims against broader evidence, the model inevitably encounters extraneous contextual information, alongside natural paraphrasing, synonym substitution, and syntactic variation. These factors naturally produce non-zero residuals even when the core facts align perfectly. To prevent the model from being overwhelmed by contextual noise or confusing benign linguistic variations with

genuine factual errors, we introduce an adaptive gating mechanism that selectively amplifies significant discrepancies while suppressing noise.

Given the residual descriptor r_k , we first project it into a lower-dimensional space of dimension d_h and compute a candidate feature vector and a gating mask in parallel:

$$\phi_k = \text{ReLU}(W_f r_k + b_f), \quad (5)$$

$$g_k = \sigma(W_g r_k + b_g), \quad (6)$$

where $W_f, W_g \in \mathbb{R}^{d_h \times 2d}$ are learnable projection matrices, $b_f, b_g \in \mathbb{R}^{d_h}$ are bias terms, and σ is the sigmoid function. The gated output is then obtained via element-wise multiplication followed by layer normalization:

$$\tilde{r}_k = \text{LN}(\phi_k \odot g_k). \quad (7)$$

The gating mechanism allows the network to learn dimension-wise filtering: for dimensions where the residual reflects extraneous context or a harmless paraphrase, g_k approaches zero, effectively suppressing the signal; for dimensions where the residual indicates a critical factual deviation, g_k approaches one, allowing the discrepancy to pass through and be amplified by the non-linear transform ϕ_k . By resolving this contextual noise dynamically within the embedding space, the gating mechanism enables highly sensitive verification directly over broad evidence documents.

4.4 Aggregation and Scoring

After the GRD operator processes all K slots, we obtain a set of gated residual features $\{\tilde{r}_1, \dots, \tilde{r}_K\}$. In factuality verification, even a single decisive factual discrepancy can render an otherwise largely correct statement unsupported. We therefore adopt element-wise max pooling across slots to extract the aggregated feature $\bar{r} \in \mathbb{R}^{d_h}$, where its j -th element is computed as:

$$\bar{r}_j = \max_{k=1}^K \tilde{r}_{k,j}. \quad (8)$$

This operation ensures that the strongest local contradiction signal is preserved and not averaged away by other semantically consistent slots.

Finally, the aggregated feature is passed through a two-layer MLP with a sigmoid output σ to produce the verification score:

$$\hat{y} = \sigma(W_2 \text{ReLU}(W_1 \bar{r} + b_1) + b_2), \quad (9)$$

where W_1 and W_2 are the learnable weight matrices of the hidden and output layers, b_1 and b_2 are the corresponding bias terms, and $\hat{y} \in [0, 1]$ represents the final predicted probability that the evidence entails the claim.

5 Experiments

5.1 Settings

Datasets. We evaluate our framework on eight diverse datasets spanning factuality verification and natural language inference: FEVER (Thorne et al., 2018), SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), WebNLG (Gardent et al., 2017), Wiki-EN-Sentence (Wiki-EN) (Liu et al., 2025), Wiki-Bio-Hallu (Bio-S/H) (Liu et al., 2025), Climate-FEVER (Climate) (Diggelmann et al., 2020), and SciFact (Wadden et al., 2020). All evaluations are formulated as binary classification tasks with supported and unsupported labels. For datasets with standard NLI formats, we map the original three-way labels to this binary setup by treating entailment as supported, while consolidating contradiction and neutral into unsupported. Detailed descriptions and statistics for all datasets are provided in Appendix B.

Metrics. We evaluate all models across two dimensions: verification performance and inference efficiency. For performance, we report the macro-averaged F1 score (Macro-F1) across all benchmarks to ensure a balanced assessment. For efficiency, we report inference latency (GPU time) and theoretical computational complexity (FLOPs).

Baselines. We compare our framework against three categories of baseline methods to comprehensively evaluate its effectiveness and efficiency. **LLM-based methods** serve as high-performance zero-shot upper bounds. We evaluate the instruction-tuned Qwen2.5 series (Yang et al., 2024) and Llama3.2 (Grattafiori et al., 2024) to establish a performance reference. **NLI-based cross-encoders** represent the dominant pairwise verification paradigm and serve as our primary competitive baselines. We include MiniCheck-DeBERTa (specifically optimized for fact-checking) (Tang et al., 2024), DeBERTa-v3-NLI, and NLI-RoBERTa (Liu et al., 2019). **Embedding-based methods** serve as direct architectural counterparts. We evaluate BERTScore (Zhang et al., 2019) and ColBERT (Khattab and Zaharia, 2020) as symmetric similarity baselines, alongside E-Verify (Liu et al.,

2025), a representative embedding-based baseline for factuality verification. Detailed descriptions, architectural properties, and implementation sources for all baseline models are provided in Appendix C.

5.2 Verification Performance

We first evaluate the discriminative accuracy of all models across eight datasets. Generative LLMs are evaluated under a zero-shot prompting setting to gauge their inherent reasoning capabilities, while other baselines (NLI and embedding-based) perform direct inference based on their standard paradigms. Table 1 presents the overall Macro-F1 results. We organize our analysis around five key findings that highlight the specific capabilities and application boundaries of the GRD framework.

Symmetric metrics fail at factuality verification. Standard embedding-based methods like BERTScore (33.00%) and ColBERT (36.65%) perform poorly. Despite ColBERT’s fine-grained late interaction, its reliance on symmetric cosine similarity forces it to assign high scores to unsupported claims due to substantial lexical overlap. This confirms our foundational premise: factuality verification inherently requires explicit *asymmetric entailment* modeling to distinguish subtle factual contradictions from benign paraphrasing.

GRD closes the bi-encoder vs. cross-encoder gap. By explicitly modeling this asymmetry via dual-view residual extraction, GRD achieves an average Macro-F1 of 79.99%, outperforming the previous state-of-the-art bi-encoder, E-Verify, by 10.04 points. Furthermore, GRD successfully narrows the performance gap with the best cross-encoder (DeBERTa-v3-NLI, 81.14%) to just 1.15 points, while operating strictly within the efficient bi-encoder paradigm. On specific benchmarks such as SNLI (96.17%), Wiki-EN (98.15%), and Climate-FEVER (64.54%), GRD achieves the best results among all non-LLM baselines, demonstrating the discriminative strength of our difference-dominant approach.

Exceptional robustness to long evidence. A critical limitation of prior bi-encoders like E-Verify is their reliance on pre-atomized evidence to avoid noise masking. To systematically evaluate this, we concatenate evidence sentences in Wiki-Bio-Hallu into passages of increasing length (100–2,000 characters). As illustrated in Figure 3, E-Verify suffers severe degradation as context grows: on Bio-S, its Macro-F1 collapses from 79.15% (atom-level) to below 30% (2,000 chars). In contrast, GRD main-

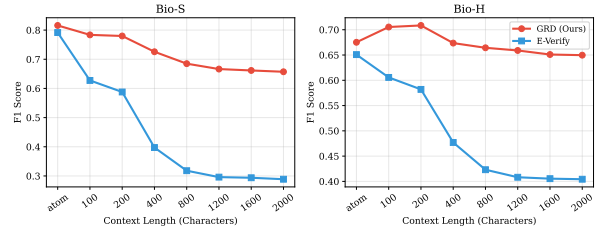


Figure 3: Context-length robustness on Bio-S/H. Evidence sentences are concatenated into passages of increasing character length. GRD maintains stable performance while E-Verify degrades sharply as context grows.

tains stable performance across all context lengths. On Bio-H, GRD even benefits from moderate context inclusion, peaking above 70% for 100–200-character passages compared with 67.52% at the atomic level. This strongly confirms that our adaptive gating mechanism successfully filters contextual noise, enabling direct and robust verification against lengthy documents.

Architectural efficiency vs. LLM scaling. While the 7B-parameter Qwen model establishes an upper bound (88.45%) due to its massive capacity for complex reasoning (especially on reasoning-heavy datasets like SciFact), this zero-shot capability does not scale down efficiently. Smaller generative models, such as Llama-3.2-3B (67.05%) and Qwen2.5-0.5B (74.01%), fall significantly behind GRD despite having vastly more parameters. This demonstrates that for factuality verification, a purpose-built discriminative architecture with targeted asymmetric modeling is far more effective and computationally viable than relying on the general-purpose capabilities of smaller LLMs.

Limitations on reasoning-heavy benchmarks. While GRD delivers stable performance across benchmarks where factual matching is the primary signal, both bi-encoders and cross-encoders lag behind massive LLMs on Climate-FEVER and SciFact. SciFact, in particular, demands complex scientific reasoning and domain knowledge beyond direct semantic entailment, inherently favoring the zero-shot inference capabilities of large-scale generative models. This highlights a natural boundary: embedding-based approaches like GRD are optimal for high-throughput textual discrepancy detection, whereas multi-step logical inference remains the domain of capacity-heavy LLMs.

Impact of Task-Specific Fine-Tuning. To ensure that the performance gains of GRD stem from its architectural design rather than merely the advan-

Model	FEVER	SNLI	MNLI	Wiki-EN	Bio-S	Bio-H	WebNLG	Climate	SciFact	Avg.
<i>LLM-based</i>										
Qwen2.5-7B [†]	93.51	95.76	92.69	98.25	93.15	82.92	86.41	69.84	83.54	88.45
Llama-3.2-3B	89.59	91.73	82.44	89.07	80.82	51.91	57.97	32.79	27.14	67.05
Qwen2.5-0.5B	82.43	89.30	78.55	86.05	72.75	66.74	77.54	54.47	58.27	74.01
<i>NLI-based</i>										
MiniCheck-DeBERTa	<u>89.72</u>	90.68	84.83	91.55	<u>87.16</u>	76.95	59.67	48.96	52.50	75.78
DeBERTa-v3-NLI	93.46	<u>94.23</u>	92.93	<u>96.57</u>	88.30	<u>74.68</u>	75.65	53.68	60.73	81.14
NLI-RoBERTa	87.23	<u>94.87</u>	<u>90.72</u>	93.34	78.48	63.52	61.52	<u>55.97</u>	54.26	75.55
<i>Embedding-based</i>										
BERTScore	33.56	33.77	34.21	36.61	37.87	25.35	23.77	32.14	39.69	33.00
ColBERT	33.60	42.39	35.52	53.08	38.00	27.20	28.24	32.14	39.69	36.65
E-Verify	60.87	81.23	79.22	<u>96.97</u>	79.15	65.08	65.29	52.81	48.97	69.95
GRD (Ours)	87.42	96.17	<u>90.43</u>	98.15	81.59	70.85	<u>72.05</u>	64.54	<u>58.69</u>	<u>79.99</u>

Table 1: Factuality verification results (Macro-F1 %) across benchmarks. GRD outperforms its bi-encoder counterpart, E-Verify, by 10.04 points. Crucially, GRD narrows the performance gap with computationally expensive cross-encoders (DeBERTa-v3-NLI) to just 1.15 points, while consistently outperforming small-parameter LLMs (e.g., Llama-3.2-3B and Qwen2.5-0.5B). [†]Results serve as an upper-bound reference.

tage of task-specific training, we conducted controlled fine-tuning experiments. As detailed in Appendix E, we fine-tuned both cross-encoder and bi-encoder baselines on the exact same training data as GRD. Under matched training conditions, GRD continues to outperform the capacity-matched embedding baseline (E-Verify) by over 7 absolute points. Furthermore, GRD demonstrates superior transferability on structurally distinct datasets (e.g., WebNLG and MNLI) compared to from-scratch fine-tuned baselines, confirming that its difference-dominant architecture learns robust structural discrepancy detection rather than merely memorizing in-domain distributions.

5.3 Inference Efficiency and Scalability

We evaluate inference efficiency in a realistic retrieval-augmented scenario using the FEVER dataset. For 5,000 queries, BM25 retrieves the top-5 evidence candidates, and the model verifies the claim against all candidates. Latency measurements are conducted on a RTX 4090 GPU with torch.compile optimization (vLLM for generative models), and all reported latencies are averaged over 5,000 claims. Crucially, cross-encoder architectures (LLMs and NLI models) require redundant joint encoding for every evidence–claim pair. In contrast, bi-encoders like GRD decouple this process: the lengthy evidence is encoded offline just once, leaving only the lightweight claim encoding and embedding-space interaction to be computed online. In Table 2, we report parameters, theoretical FLOPs, and end-to-end latency per claim.

Orders-of-Magnitude Reduction in Online Cost.

As shown in Table 2, GRD reduces the online cost by over an order of magnitude compared to the strongest cross-encoder (DeBERTa-v3-NLI), dropping from 83.52 GFLOPs to just 3.96 GFLOPs per claim. This translates to a $23.55\times$ end-to-end speedup (0.33ms vs. 7.77ms latency). While E-Verify exhibits a higher speedup due to its smaller backbone, it sacrifices over 10 points in accuracy (as shown in Table 1). Overall, GRD approaches cross-encoder-level verification accuracy while retaining bi-encoder-level efficiency.

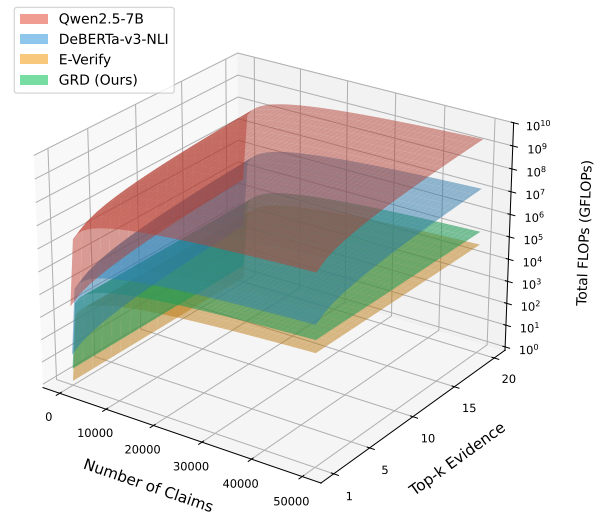


Figure 4: Total computational FLOPs scaling on the FEVER dataset with respect to the number of claims and top- k retrieved evidence. Joint-encoding architectures exhibit sharp polynomial growth, whereas GRD maintains orders-of-magnitude lower cost along the claim axis and negligible marginal cost along the top- k axis.

Efficient Scaling Across Claims and Evidence.

Using the average token lengths from the FEVER

Model	Avg Tokens	Params	FLOPs-Offline	FLOPs-Online	Latency	Speedup
<i>LLM-based</i>						
Qwen2.5-7B	157.3	7.7B	–	12.09T	50.32ms	0.15×
Llama-3.2-3B	153.8	3.2B	–	1.10T	21.33ms	0.36×
Qwen2.5-0.5B	157.3	0.5B	–	828.77G	2.20ms	3.53×
<i>NLI-based</i>						
MiniCheck-DeBERTa	97.8	435M	–	291.82G	23.59ms	0.33×
DeBERTa-v3-NLI	97.8	184M	–	83.52G	7.77ms	1.0×
NLI-RoBERTa	112.5	125M	–	47.83G	4.16ms	1.87×
<i>Embedding-based</i>						
E-Verify	94.2 / 12.7	114M	8.13G	1.11G	0.13ms	59.77×
GRD (Ours)	94.2 / 12.7	342M	28.69G	3.96G	0.33ms	23.55×

Table 2: Efficiency comparison on the FEVER dataset. Avg Tokens indicates the concatenated sequence length ($E + C$) for joint-encoding models, and the individual lengths of evidence and claim (E/C) for embedding-based models. Reported FLOPs and latency denote the average computational cost per claim, evaluated against the top-5 evidence candidates retrieved by BM25 (averaged across 5,000 queries). By decoupling evidence and claim encoding, GRD reduces online computational cost by over an order of magnitude compared to cross-encoders, achieving a 23.55× speedup over DeBERTa-v3-NLI while maintaining competitive accuracy.

scenario, Figure 4 visualizes the total theoretical FLOPs as the number of claims and retrieved evidence candidates (top- k) increase. For joint-encoding models (Qwen2.5-7B and DeBERTa-v3-NLI), the computational cost increases sharply along both axes, as every verification requires a full forward pass over the concatenated evidence and claim sequence. In contrast, GRD demonstrates substantially lower growth slopes on both fronts. Along the claim axis, GRD encodes only the short claim text online (averaging 12.7 tokens), bypassing the repetitive encoding of the lengthy evidence context. This keeps the base computational footprint orders of magnitude lower than cross-encoders. Along the top- k axis, the cost surface is nearly flat. Since the claim is encoded exactly once and evidence embeddings are fetched from an offline index, scaling from 1 to 20 candidates merely adds lightweight embedding-space operations, yielding negligible marginal cost.

5.4 Ablation Study

Variant	FEVER	MNLI	Bio-S	WebNLG
GRD	87.42	90.43	81.59	72.05
w/o directional residual	82.82	87.35	67.83	60.19
w/o magnitude residual	81.13	88.52	72.63	69.04
w/o gating amplification	85.20	84.49	80.08	70.54

Table 3: Ablation study on GRD components (Macro-F1 %). Excluding the directional residual, magnitude residual, or adaptive gating leads to significant performance degradation across all benchmarks, demonstrating their complementary roles in robustly capturing asymmetric factual discrepancies.

Table 3 examines the contribution of each com-

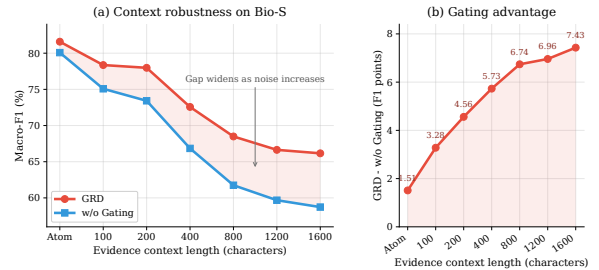


Figure 5: Context-length stress test on the Bio-S dataset. (a) compares the Macro-F1 of the full GRD model against the variant without gating across increasing evidence context lengths. (b) plots the absolute performance gap between the two models. The steadily widening gap demonstrates the gate’s effectiveness in filtering accumulating contextual noise.

ponent in the GRD framework. Removing the directional residual δ_k causes severe degradation on generation-based hallucination benchmarks, dropping by 13.76 absolute points on Bio-S and 11.86 points on WebNLG, which confirms its necessity for capturing explicit asymmetric entailment. Meanwhile, excluding the magnitude residual μ_k yields significant drops on Bio-S (−8.96%) and FEVER (−6.29%), proving its role in gauging the severity of factual contradictions; notably, it is even more impactful than the directional view on FEVER.

Finally, removing the adaptive gating mechanism g_k consistently hurts performance across all benchmarks. To explicitly verify its role as a contextual noise filter, we conduct a context-length stress test on the Bio-S dataset. By systematically increasing the evidence length, we introduce more irrelevant background information. As illustrated

in Figure 5(a), the performance of the variant without gating degrades rapidly as context expands. Figure 5(b) further highlights that the performance gap between the full GRD and the ablated variant steadily widens, growing from 1.51 points at the atomic level to 7.43 points at 1,600 characters. This growing delta quantitatively confirms that the gating mechanism successfully suppresses the accumulating contextual noise, preventing the model from being overwhelmed by harmless lexical overlap in long documents. Furthermore, the steep performance drop on MNLI (−5.94% without gating in Table 3) corroborates this finding, showing that dynamic gating is equally essential for filtering diverse syntactic variations.

Overall, these results demonstrate that all three components function complementarily to ensure robust and sensitive verification. Detailed hyperparameter analysis regarding the encoder backbone capacity and the number of semantic slots K is provided in Appendix D.

6 Conclusion

In this paper, we introduced the Gated Residual Discrepancy (GRD) framework to address the efficiency and scalability bottlenecks in factuality verification. By explicitly modeling asymmetric semantic entailment through dual-view residuals and adaptive gating, GRD successfully bridges the performance gap between lightweight bi-encoders and computationally intensive cross-encoders. Crucially, GRD decouples the verification pipeline into offline evidence indexing and lightweight online interaction. This architectural design reduces online computational costs by over one order of magnitude compared to standard cross-encoders. Furthermore, it exhibits negligible marginal cost when scaling the number of retrieved evidence candidates, making it suitable for large-scale, high-throughput verification systems. While GRD excels at identifying direct textual discrepancies and generation-based hallucinations, a natural limitation is its performance on reasoning-heavy tasks. Such tasks inherently demand the multi-step logical inference and extensive world knowledge currently dominated by massive large language models. Therefore, future work will explore hybrid, cascading verification pipelines: deploying GRD as a highly efficient and robust first-pass filter to resolve the vast majority of factual queries, while selectively routing only the most complex, reasoning-intensive

claims to capacity-heavy LLMs or cross-encoders.

Limitations

While the Gated Residual Discrepancy (GRD) framework provides a highly efficient bi-encoder solution for factuality verification, its architectural design imposes inherent boundaries on its capabilities. GRD is explicitly optimized for identifying direct textual discrepancies and generation-based hallucinations, where factual consistency can be reliably assessed through geometric vector differences and semantic alignment. Consequently, the framework underperforms on reasoning-heavy tasks, such as verifying complex scientific claims (e.g., the SciFact benchmark). Because embedding-based methods compress text into dense vector spaces, they are generally better at capturing direct semantic shifts than performing multi-step logical deduction. For claims that require complex inference chains, implicit knowledge synthesis, or deep domain-specific reasoning, the late-interaction and residual-based mechanisms of GRD are insufficient compared to the deep cross-attention layers of massive language models. Therefore, in real-world deployments, GRD is best positioned as a high-throughput, first-pass filter for explicit factual errors, rather than a standalone reasoning engine for highly complex logical tasks.

Acknowledgments

This work was supported by the National Science and Technology Major Project under Grant 2022ZD0120204.

The authors used generative AI tools solely for language polishing and improving the clarity of the manuscript. All technical content, research ideas, analyses, and conclusions were developed and verified by the authors.

References

- Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 670–680.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bu-
lian, Massimiliano Ciaramita, and Markus Leip-
pold. 2020. Climate-fever: A dataset for verifica-
tion of real-world climate claims. *arXiv preprint arXiv:2012.00614*.
- Florian E Dörner, Vivian Y Nastl, and Moritz Hardt. 2024. Limits to scalable evaluation at the fron-
tier: Llm as judge won’t beat twice the data. *arXiv preprint arXiv:2410.13341*.
- Wenbiao Du, Xiuqi Yang, Zeyang Liu, Boya Zhang, and Jingfeng Xue. 2026. Evolva: A multi-turn contex-
tual attack for long-reasoning llms. In *International Conference on Knowledge Science, Engineering and Management*, pages 617–631. Springer.
- Jiarun Fu, Lizhong Ding, Qiuning Wei, Yuhan Guo, Yurong Cheng, and Junyu Zhang. 2026. Counterfac-
tual planning for generalizable agents’ actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29432–29440.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg chal-
lenge: Generating text from rdf data. In *Proceedings of the 10th international conference on natural lan-
guage generation*, pages 124–133.
- Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 con-
ference on empirical methods in natural language processing*, pages 7765–7784.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-
Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of mod-
els. *arXiv preprint arXiv:2407.21783*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adap-
tation of large language models. *arXiv preprint arXiv:2106.09685*.
- Hui Huang, Muyun Yang, and Yuki Arase. 2026. Diva: Fine-grained factuality verification with
agentic-discriminative verifier. *arXiv preprint arXiv:2601.03605*.
- Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. Openfactcheck: A unified framework
for factuality evaluation of llms. *arXiv:2408.11832*.
- Rui Jiao, Yue Zhang, and Jinku Li. 2025. Trustworthy reasoning: Evaluating and enhancing factual accu-
racy in llm intermediate thought processes. *arXiv preprint arXiv:2507.22940*.
- Omar Khattab and Matei Zaharia. 2020. Colbert: Effi-
cient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.
- Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-
based models for inconsistency detection in summa-
rization. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Ko-
siolek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based
permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–
3753. PMLR.
- Zihan Liao, Hang Yu, Jianguo Li, Jun Wang, and Wei Zhang. 2024. D2llm: Decomposed and dis-
tilled large language models for semantic search. *arXiv:2406.17262*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man-
dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining ap-
proach. *arXiv preprint arXiv:1907.11692*.
- Zeyang Liu, Jingfeng Xue, Xiuqi Yang, Wenbiao Du, Jiarun Fu, Junbao Chen, Wenjie Guo, and Yong Wang. 2025. E-verify: A paradigm shift to scalable
embedding-based factuality verification. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5759–5776.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation
of factual precision in long form text generation. *arXiv:2305.14251*.
- Jihwan Oh, Jeonghwan Choi, Nicole Hee-Yoen Kim, Taewon Yun, and Hwanjun Song. 2025. Learning to
verify summary facts with fine-grained llm feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 230–242.
- Qian Pan, Zahra Ashktorab, Michael Desmond, Mar-
tin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-
centered design recommendations for llm-as-a-judge. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 16–29.
- Subhey Sadi Rahman, Md Adnanul Islam, Md Mahbub Alam, Musarrat Zeba, Md Abdur Rahman, Sadia Sul-
tana Chowha, Mohaimenul Azam Khan Raiaan, and

- Sami Azam. 2026. Hallucination to truth: a review of fact-checking and factuality evaluation in large language models. *Artificial Intelligence Review*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pages 3982–3992.
- Giovanni Servadio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. 2025. Are the hidden states hiding something? testing the limits of factuality-encoding capabilities in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6089–6104.
- Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. 2024. Atomic inference for nli with generated facts as atoms. In *EMNLP*.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *EMNLP*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv:2412.15115*.
- Xiuqi Yang, Xiaojun Xu, Wenbiao Du, Zeyang Liu, Junbao Chen, Yifeng Fu, and Jingjing Hu. 2026. Diffhgcl: Task-aware diffusion-augmented heterogeneous graph contrastive learning for web api recommendation. *Expert Systems with Applications*, page 133349.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv:1904.09675*.
- Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523.

A Training Details

A.1 Training Objective

The model is trained with binary cross-entropy loss on evidence–claim pairs labeled as supported ($y = 1$) or unsupported ($y = 0$). The overall training objective is formulated as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] . \quad (10)$$

A.2 Two-Stage Training Strategy

We adopt a two-stage training strategy. In the first stage, the pre-trained language model is frozen and only the PMA and GRD parameters are optimized, allowing the verification modules to learn effective difference interaction patterns on stable representations. In the second stage, we introduce Low-Rank Adaptation (LoRA) (Hu et al., 2021) to fine-tune the language model, enabling the encoder to adapt its representations to better serve the difference-dominant verification paradigm. To preserve the general linguistic knowledge learned during pre-training, LoRA is applied only to the last two Transformer layers of the encoder, restricting adaptation to the highest-level semantic representations while keeping the lower layers intact. This staged strategy improves optimization stability compared with jointly training newly initialized interaction layers and adapting the encoder from the beginning.

A.3 Environment and Hyperparameters

We use BGE-large-en-v1.5 (Xiao et al., 2024) as the pre-trained encoder backbone, which produces

1,024-dimensional token embeddings from a 24-layer Transformer, and train the model on the training set of FEVER, Wiki-EN-Sentence, MNLI, SNLI and WANLI. The PMA module uses $K=8$ seed vectors with 8 attention heads, and the GRD verifier has a hidden dimension $d_h=512$. Dropout is set to 0.1 throughout. In Stage 1, the PMA and GRD parameters are optimized with Adam (learning rate 5×10^{-5}) for 4 epochs. In Stage 2, LoRA adapters with rank $r=8$ are injected into $\mathbf{W}_Q$ and $\mathbf{W}_V$ of the last two encoder layers, and the entire model is trained with a reduced learning rate of 2×10^{-5} for another 4 epochs. The batch size is 32 in both stages. All experiments are conducted on a single NVIDIA RTX 4090 GPU (24 GB) with an AMD Ryzen 9950X CPU and 128 GB RAM.

B Dataset Details

Table 4 summarizes the detailed descriptions and statistics of the eight datasets used for evaluation.

C Baseline Details

Table 5 provides a comprehensive overview of the baseline methods evaluated in our experiments, including their architectural paradigms, specific descriptions, and implementation sources.

D Hyperparameter Analysis

Encoder Backbone. Figure 6 compares four encoder backbones of varying capacities. Performance generally scales with model size: upgrading from the lightweight all-MiniLM-L6-v2 (22.7M parameters) to BGE-large-en-v1.5 (335M) yields steady improvements, gaining 3.75 absolute points on FEVER and 4.26 points on MNLI. Notably, the 22.7M MiniLM backbone exhibits surprisingly strong baseline performance, even marginally outperforming the 33M BGE-small on FEVER and Bio-S. This indicates that the GRD interaction module can effectively extract and amplify factual discrepancies even from highly constrained representations. We adopt BGE-large-en-v1.5 as the default backbone due to its optimal overall accuracy.

Number of Semantic Slots. Figure 7 investigates the effect of the number of semantic slots K produced by the PMA module. Increasing K from 1 to 8 provides substantial gains on document-level verification tasks, improving FEVER by 1.75 points and Bio-S by 1.43 points. However, the optimal K varies across datasets: while WebNLG and MNLI peak earlier at $K = 2$ and $K = 4$

respectively, further increasing K to 16 causes notable degradation across most benchmarks (e.g., -1.77 points on FEVER and -1.73 points on WebNLG). This degradation likely stems from the over-fragmentation of semantic information, which dilutes the discriminative power of individual slot vectors. We select $K = 8$ as the default setting, achieving the best overall trade-off between fine-grained representational capacity and verification robustness.

E Controlled Fine-Tuning Comparisons

To address potential concerns that GRD’s performance advantage might solely arise from task-specific training on the FEVER and Wiki-EN-Sentence datasets, we evaluate baseline models under rigorously matched training conditions. Table 6 presents the macro-F1 scores of DeBERTa-v3 and E-Verify when evaluated off-the-shelf, continued fine-tuned, and fine-tuned from scratch using the identical training data as GRD.

Even when E-Verify is fine-tuned (FT) and its backbone is upgraded to match the parameter capacity of GRD (FT on Large Backbone), GRD still outperforms it by over 7 absolute points on average. Additionally, the from-scratch fine-tuned DeBERTa-v3-large model struggles to generalize to out-of-domain datasets with different syntactic structures, such as WebNLG (63.26%) and MNLI (82.71%). In contrast, GRD achieves 72.05% and 90.43% on these datasets, respectively. This demonstrates that GRD’s difference-dominant residual architecture provides a strong inductive bias for detecting structural and local mismatches, driving its superior cross-dataset generalization.

Dataset	Description	#Pairs
FEVER (Thorne et al., 2018)	The NLI-formatted version of the FEVER benchmark, where claims are generated from Wikipedia sentences and verified against the original passages.	12,656
SNLI (Bowman et al., 2015)	A large-scale NLI corpus where premises are drawn from Flickr image captions and hypotheses are crowd-sourced.	6,605
MNLI (Williams et al., 2018)	A multi-genre NLI corpus spanning ten domains (e.g., fiction, government, telephone), designed to test cross-domain generalization of textual entailment models.	6,692
Wiki-EN-Sentence (Liu et al., 2025)	A sentence-level factual verification dataset constructed from Wikipedia, where each claim is paired with a single evidence sentence.	5,000
WebNLG (Gardent et al., 2017)	An NLI dataset derived from the WebNLG challenge, where premises are natural language descriptions of structured knowledge graph triples and hypotheses are generated text.	48,869
Wiki-Bio-Hallu (Liu et al., 2025)	A hallucination detection dataset for biography generation. Each generated biography is verified against its Wikipedia source. It includes a <i>simple</i> subset with controlled factual errors in numbers, dates, or entities, and a <i>hard</i> subset where errors naturally occur in LLM-generated text.	476
Climate-FEVER (Diggelmann et al., 2020)	A climate science fact verification dataset where claims about climate change are verified against evidence from scientific sources.	1,381
SciFact (Wadden et al., 2020)	A scientific fact verification dataset where claims are derived from scientific abstracts and verified against the original papers.	693

Table 4: Summary of evaluation datasets. #Pairs denotes the number of evaluation pairs.

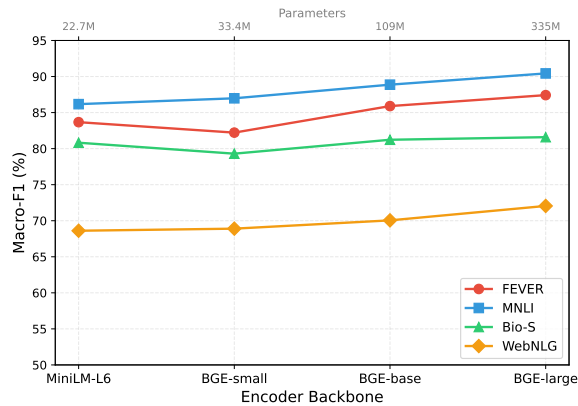


Figure 6: Effect of encoder backbone. Parameter counts are shown on the top axis.

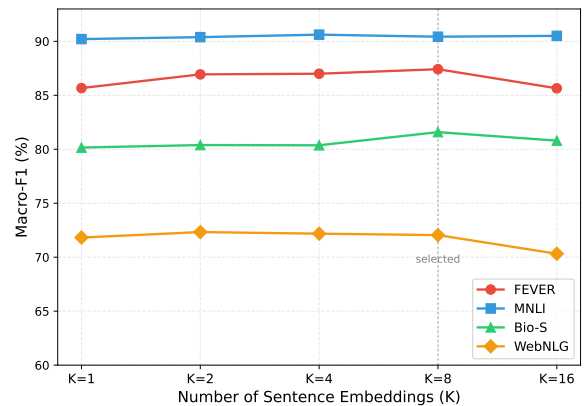


Figure 7: Effect of the number of semantic slots K .

Category	Model	Description
LLM-based	Qwen2.5 series (Yang et al., 2024)	A family of instruction-tuned language models from Alibaba, including 0.5B and 7B variants. We evaluate these models to establish a high-performance zero-shot upper bound for factuality verification. https://huggingface.co/collections/Qwen/qwen25
	Llama3.2-3B (Grattafiori et al., 2024)	A lightweight, instruction-tuned large language model from Meta, optimized for efficient inference. We include it to assess the zero-shot reasoning capabilities of compact generative models on verification tasks. https://huggingface.co/meta-llama/Llama-3.2-3B
NLI-based	MiniCheck-DeBERTa (Tang et al., 2024)	A DeBERTa-v3-large model specifically trained for grounded fact-checking tasks. Optimized for detecting factual inconsistencies in generated text against reference documents. https://huggingface.co/lytang/MiniCheck-DeBERTa-v3-Large
	DeBERTa-v3-NLI	A cross-encoder based on DeBERTa-v3-base fine-tuned on MNLI, FEVER, and ANLI datasets. Employs disentangled attention and an enhanced mask decoder for improved NLI performance. https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli
	NLI-RoBERTa (Liu et al., 2019)	A RoBERTa-base cross-encoder fine-tuned on multiple NLI benchmarks. Provides a lightweight alternative to DeBERTa-based models with competitive accuracy. https://huggingface.co/cross-encoder/nli-roberta-base
Embedding-based	BERTScore (Zhang et al., 2019)	A widely-used automatic evaluation metric that computes token-level contextual embedding similarity between reference and candidate texts. It serves as a symmetric similarity baseline for factuality verification. https://github.com/Tiiiger/bert_score
	ColBERT (Khattab and Zaharia, 2020)	A late-interaction retrieval model that independently encodes queries and documents into token-level embeddings, then performs fine-grained MaxSim-based matching. Despite its token-level interaction design, it relies on symmetric cosine similarity. https://huggingface.co/colbert-ir/colbertv2.0
	E-Verify (Liu et al., 2025)	A pioneering bi-encoder framework for embedding-level factuality verification. Employs multi-vector representations and fuses multiple features including pairwise similarity and discrepancy features. https://github.com/lzy248/e-verify
	GRD (Ours)	Our proposed bi-encoder method. It adopts a difference-dominant paradigm that extracts dual-view residuals from multi-vector representations, then applies adaptive gating to amplify critical factual deviations while suppressing benign paraphrasing variations.

Table 5: Overview of factuality verification methods across three paradigms.

Model	FEVER	SNLI	MNLI	Wiki-EN	Bio-S	Bio-H	WebNLG	Climate	SciFact	Avg.
DeBERTa-v3-NLI (Off-the-shelf)	93.46	94.23	92.93	96.57	88.30	74.68	75.65	53.68	60.73	81.14
DeBERTa-v3-NLI (Continue FT)	92.34	94.65	93.29	98.59	88.75	75.75	75.38	59.35	60.85	82.11
DeBERTa-v3-large (Scratch FT)	91.82	84.67	82.71	98.16	67.67	63.00	63.26	41.10	42.34	70.53
E-Verify (Off-the-shelf)	60.87	81.23	79.22	96.97	79.15	65.08	65.29	52.81	48.97	69.95
E-Verify (FT)	78.69	84.68	80.56	96.98	79.22	64.87	66.04	53.90	51.56	72.95
E-Verify (FT on Large Backbone)	81.94	83.83	81.91	94.58	75.48	65.18	66.93	53.75	50.61	72.69
GRD (Ours)	87.42	96.17	90.43	98.15	81.59	70.85	72.05	64.54	58.69	79.99

Table 6: Factuality verification results with controlled fine-tuning comparisons (Macro-F1 %). Under matched training conditions and backbone capacities, GRD maintains a significant architectural advantage over standard embedding-based methods and demonstrates superior generalization to structurally distinct benchmarks.