

Format as Nuisance: Intervention-Inspired Cross-Format Distillation for Knowledge Conflicts

Yue Yu¹, Wang Luo², Shenghao Ye³, Zepeng Bao¹,
Jiacheng Liu¹, Yongqi Li¹, Wenxin Tang⁴, Tiejun Qian^{1*}

¹Wuhan University ²Sun Yat-sen University

³University of Science and Technology of China ⁴Tsinghua University

{yuyue1, qty}@whu.edu.cn

Abstract

Large Language Models (LLMs) encode parametric knowledge but can exhibit format bias when resolving knowledge conflicts. When evidence is rendered as text, tables, or knowledge graphs, LLMs can favor certain formats despite unchanged semantic content. Existing work mainly diagnoses this bias, while existing inference-time mitigation remains tied to a single rendered prompt. We propose FaDe, a cross-format knowledge distillation framework that treats evidence format as a controllable nuisance variable and marginalizes teacher predictions over controlled format assignments. FaDe constructs a format-marginalized soft target from these renderings, then trains the student with an asymmetric two-path objective that combines heterogeneous-format debiasing with all-text anchoring. Experiments on ConflictBank and two QA-derived controlled conflict sets with four model backbones show FaDe reduces format bias while improving dual-source coverage and cross-format stability under single-pass inference. Our code is available at: <https://github.com/Kaia91/FaDe>.

1 Introduction

Large Language Models (LLMs) encode vast parametric knowledge during pretraining, yet this knowledge frequently conflicts with external evidence during inference (Longpre et al., 2021; Xie et al., 2024; Chen et al., 2022). In practice, evidence often arrives in heterogeneous formats, and models must synthesize it reliably across surface forms. Recent work shows that LLMs exhibit systematic format bias: the same factual content can lead to different preferences when rendered in different formats (Su et al., 2024; Liu et al., 2026). This format preference acts as a nuisance variable that distorts model judgments independently of evidence quality, undermining reliability in format-heterogeneous settings, as illustrated in Figure 1.

*Corresponding author.

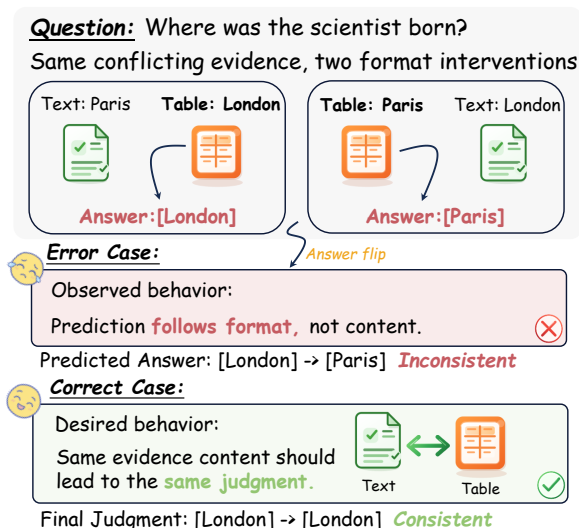


Figure 1: Counterfactual format bias: holding conflicting content fixed while swapping evidence formats can flip the model’s prediction.

We make this view explicit by treating format bias as a nuisance-variable problem. The task-relevant factor is the semantic content of the evidence, while the rendering format should not determine which conflicting claim is preferred. However, a model never observes content in isolation. Each piece of evidence is assigned to a concrete surface format before inference. Thus, standard inference observes only one rendered prompt, where content and format are already entangled. A response to this single rendering cannot reveal whether the model follows a claim because of its semantic content or because of its assigned format. This entanglement makes format bias different from ordinary input noise. The nuisance variable is discrete, observable, and directly controllable through rendering, but its effect is only reflected through the model’s response.

This perspective motivates cross-format comparisons. Since format is controllable through rendering, the same content can be evaluated under different format assignments. A recent inference-time mitigation approach rescales attention weights to

encourage more balanced use of different evidence sources (Liu et al., 2026). However, it still operates on a single rendered prompt and can only rebalance evidence use within that rendering, leaving the model’s directional format preference largely unchanged. This limitation calls for a parameter-level learning signal derived from controlled cross-format comparisons. Distillation provides a natural vehicle for such a signal because it transfers soft preferences between competing claims, but single-rendering distillation would still inherit the teacher’s format preference. We instead query a frozen teacher under multiple controlled format assignments and average its predictions into a format-marginalized soft target. In this way, the teacher is not used as an unbiased oracle under any single format, but as a probe for estimating a less format-dependent judgment across interventions.

In this paper, we propose **FaDe** (Format-Averaged Debiasing), a cross-format knowledge distillation framework for reducing format bias in knowledge conflicts. For each content pair, FaDe constructs a format-marginalized teacher target by averaging teacher predictions across an all-text control and multiple heterogeneous format assignments. It uses a disagreement-based adaptive weight, so samples with larger cross-format teacher instability receive stronger debiasing. The student is trained with an asymmetric two-path objective. The activation path exposes the student to heterogeneous inputs and matches the marginalized target, while the anchoring path uses all-text inputs to preserve confident behavior under the standard all-text format. Experiments on three knowledge-conflict benchmarks and four 8B-class backbones show that FaDe consistently reduces format bias while improving dual-source coverage and cross-format stability, achieving an average bias reduction of 34.1% across 12 model-dataset settings.

Our contributions: (1) *New Perspective.* We formulate format bias as a nuisance-variable problem with a discrete, observable, and rendering-controllable nuisance factor, explaining why single-rendering mitigation cannot remove directional format preference. (2) *Novel Framework.* We propose FaDe, a cross-format distillation framework that constructs format-marginalized targets and uses an asymmetric activation-and-anchoring objective. (3) *Strong Performance.* FaDe reduces format bias across three benchmarks and four backbones, improving dual-source coverage and cross-format stability under single-pass inference.

2 Related Work

Knowledge Conflicts and Format Bias. LLMs may ignore, over-trust, or miscalibrate under conflicting external evidence (Longpre et al., 2021; Xie et al., 2024; Chen et al., 2022; Su et al., 2024). Their judgments are also sensitive to surface factors such as position, context placement, and response length (Zheng et al., 2023; Wang et al., 2024; Liu et al., 2024), and perturbation tests expose brittle decision boundaries (Jia and Liang, 2017; Ribeiro et al., 2020; Gardner et al., 2020). Liu et al. (2026) show that equivalent content can be preferred differently across evidence formats. We therefore treat format bias as a learned surface-format prior, not merely an input-level attention artifact.

Heterogeneous Evidence. Retrieval-augmented QA complements parametric memory with external passages (Karpukhin et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021), while structured-evidence methods use table-text encoders or unified representations over text, tables, lists, and knowledge bases (Herzig et al., 2020; Yin et al., 2020; Sun et al., 2018; Oguz et al., 2022). These systems improve evidence access but treat rendering as a faithful carrier of content. FaDe instead asks whether equivalent content changes model decisions solely because its format changes.

Knowledge Distillation. Knowledge distillation transfers teacher behavior through soft targets (Hinton et al., 2015). Generation KD has explored forward KL, reverse KL, f-divergence objectives, and on-policy variants to balance mode coverage and mode seeking (Wen et al., 2023; Gu et al., 2024; Agarwal et al., 2024). These objectives improve distribution matching, but a single-rendering teacher target can still copy the teacher’s format preference. FaDe instead distills a format-marginalized target so KD becomes a debiasing mechanism.

Intervention-Inspired Debiasing. Causal methods have inspired debiasing approaches for spurious correlations in vision and NLP (Pearl, 2009; Feder et al., 2022; Tang et al., 2020; Niu et al., 2021; Nan et al., 2021; Utama et al., 2020). We borrow this intervention perspective rather than claiming full causal identification: evidence format is finite, observed, and directly assignable by rendering over a predefined support. This makes it possible to average predictions across controlled format assignments and train toward a less format-dependent judgment during student training.

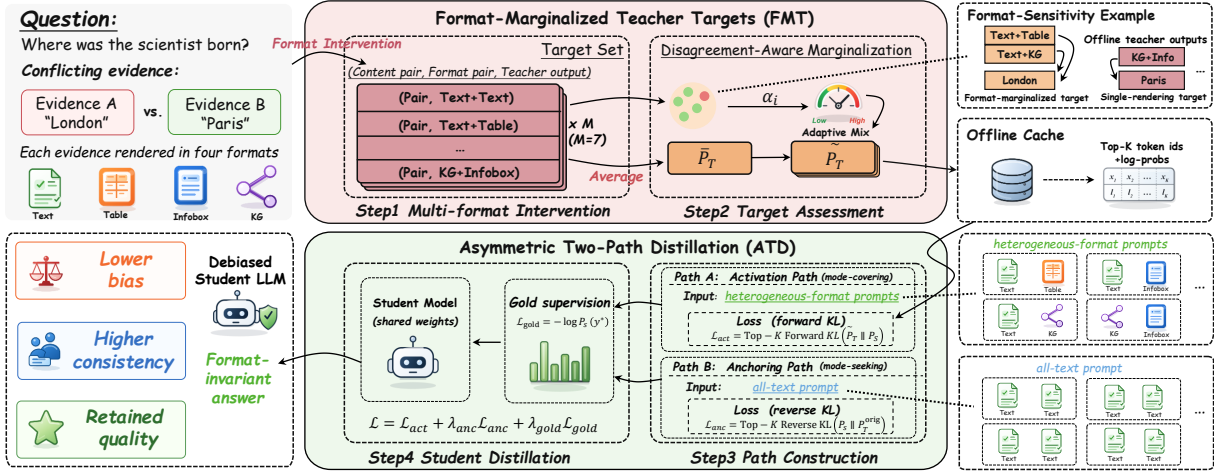


Figure 2: Overview of FaDe. For each conflict, FaDe renders the same evidence under a finite set of format assignments, marginalizes frozen-teacher predictions with disagreement-aware mixing, caches compact top- K targets, and trains a student with heterogeneous activation and all-text anchoring paths.

3 Methodology

We present **FaDe**, an intervention-inspired cross-format distillation framework for mitigating format bias in knowledge conflicts. Figure 2 summarizes the pipeline: we first model format as an assignable nuisance variable (§3.1), then construct finite-support format-marginalized teacher targets (§3.2), train with asymmetric two-path distillation (§3.3), and define the objective (§3.4).

3.1 Problem Formulation

We consider binary knowledge conflicts with two sources supporting contradictory claims. Each source has content $c \in \mathcal{C}$ and format $f \in \mathcal{F}$, with $\mathcal{F}$ covering text, table, KG, and infobox renderings. For content pair $C = (c_A, c_B)$ and assignment $F = (f_A, f_B)$, the rendered input is $X = \text{render}(C, F)$ and the output is $P(y | X)$. Format bias occurs when changing only F changes the prediction, making format a nuisance rather than task-relevant evidence.

For FaDe target construction, we use the finite intervention support

$$\mathcal{A}_{\text{eval}} = \{(\text{text}, \text{text}), (\text{text}, \text{table}), (\text{text}, \text{kg}), (\text{text}, \text{infobox}), (\text{table}, \text{kg}), (\text{table}, \text{infobox}), (\text{kg}, \text{infobox})\}.$$

This set contains one all-text control and six heterogeneous pairs ($|\mathcal{A}_{\text{eval}}| = 7$), whereas Multi-Format KD uses the full ordered heterogeneous support $\mathcal{F}_{\neq}^2 = \{(f_A, f_B) \in \mathcal{F}^2 : f_A \neq f_B\}$ with assignment-specific targets. We summarize content, format, prompt, and output with the following

lightweight causal graph

$$C \rightarrow X \leftarrow F, \quad X \rightarrow Y \quad (1)$$

where content and format jointly determine the observed prompt used to produce Y . Format cues such as table delimiters, triples, or infobox fields can therefore activate priors through $F \rightarrow X \rightarrow Y$. Because format is finite and render-assignable, we can vary it while holding content fixed. Let $x(c, a) = \text{render}(c, a)$ denote the prompt for content c under assignment a .

$$q_{\pi}^T(y | c) = \sum_{a \in \mathcal{A}_{\text{eval}}} \pi(a) \times P_T(y | x(c, a)). \quad (2)$$

With uniform $\pi(a) = 1/7$, Eq. 2 averages frozen-teacher predictions across controlled assignments for the same conflict. Because each assignment is rendered for the same content pair, the target is directly estimable over $\mathcal{A}_{\text{eval}}$ and should be read as a finite-support format-marginalized teacher target, not an unrestricted format-free causal estimand.

Assumptions and Scope. Eq. 2 assumes a controlled-rendering regime, not access to true outcome distributions. Within Eq. 1, the teacher response depends on content and format only through the rendered prompt X ; each assignment must preserve the same answer-bearing content; and the uniform π is a design choice that prevents one rendering from dominating the finite support. Under these explicit design assumptions, averaging model-conditional teacher predictions can reduce format-specific presentation priors. It can also dilute useful

signal when a format supplies information absent from other renderings, such as fields that disambiguate entities, relations, or dates. FaDe is therefore intended for semantically matched, content-preserving renderings, not settings where format changes the available evidence.

Framework Overview. FaDe implements this idea through offline teacher-target construction and student training. A frozen teacher processes each content pair under $\mathcal{A}_{\text{eval}}$; its distributions are log-averaged with a disagreement-based mixture weight, and the student learns from heterogeneous activation plus all-text anchoring signals.

3.2 Format-Marginalized Teacher Target Construction

Multi-Format Intervention. For each sample, we render $a_j \in \mathcal{A}_{\text{eval}}$ for $j = 1, \dots, M$ with $M = 7$, and obtain teacher distribution $P_T(y \mid c, a_j)$. Since content is fixed, variation across these outputs estimates sample-level format dependence.

Log-Space Marginalization. Writing $P_T^{(j)} = P_T(\cdot \mid c, a_j)$, the marginalized target is

$$\bar{P}_T = \text{Normalize} \left(\frac{1}{M} \sum_{j=1}^M P_T^{(j)} \right) \quad (3)$$

To avoid storing all full-vocabulary tensors during cache construction, we update the average incrementally in log-probability space:

$$\log \bar{P}_T^{(j)} = \text{LSE} \left(\log \bar{P}_T^{(j-1)} + \log \frac{j-1}{j}, \log P_T^{(j)} + \log \frac{1}{j} \right) \quad (4)$$

where LSE denotes LogSumExp. We measure sample-level format sensitivity by the average JS divergence between the all-text teacher prediction and each heterogeneous variant:

$$d_i = \frac{1}{|\mathcal{A}_{\text{het}}|} \sum_{a_j \in \mathcal{A}_{\text{het}}} \delta_{ij}, \quad \delta_{ij} = D_{\text{JS}} \left(Q_i^{\text{text}} \parallel Q_i^{(j)} \right) \quad (5)$$

where $Q_i^{\text{text}} = P_T(\cdot \mid c_i, a_{\text{text}})$, $Q_i^{(j)} = P_T(\cdot \mid c_i, a_j)$, $a_{\text{text}} = (\text{text}, \text{text})$, and $\mathcal{A}_{\text{het}} = \mathcal{A}_{\text{eval}} \setminus \{a_{\text{text}}\}$ contains the six heterogeneous evaluation pairs. The score d_i is mapped to a clipped linear weight α_i , yielding

$$\tilde{P}_{T,i} = \alpha_i \cdot \bar{P}_{T,i} + (1 - \alpha_i) \cdot P_{T,i}^{\text{orig}} \quad (6)$$

High-disagreement samples rely more on the marginalized target, while stable samples remain close to the original teacher, concentrating debiasing where format changes the teacher distribution. We keep the top- K tokens and renormalized log-probabilities as compact offline targets.

3.3 Asymmetric Two-Path Distillation

The student uses two paths with different input formats and divergence directions, separating debiasing from all-text anchoring. The heterogeneous path covers the marginalized teacher consensus, while the text path reduces drift from confident behavior on ordinary inputs.

Activation Path. For heterogeneous activation inputs, the student matches the marginalized teacher target with top- K forward KL:

$$\mathcal{L}_{\text{act}} = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \sum_{k=1}^K \tilde{P}_T(v_k^t) \times \left[\log \tilde{P}_T(v_k^t) - \log P_S(v_k^t) \right] \quad (7)$$

Here $\mathcal{R}$ is the response-token set, v_k^t is the k -th teacher top token at position t from the cached target, and P_S is the student distribution. Forward KL is mode-covering, assigning mass to tokens that remain plausible under the cross-format consensus rather than one teacher rendering.

Anchoring Path. For all-text anchoring inputs, the student matches the original teacher’s confident distribution through top- K reverse KL:

$$\mathcal{L}_{\text{anc}} = \frac{1}{|\mathcal{R}|} \sum_{t \in \mathcal{R}} \sum_{k=1}^K P_S(u_k^t) \times \left[\log P_S(u_k^t) - \log P_T^{\text{orig}}(u_k^t) \right] \quad (8)$$

Here u_k^t is the k -th student top token used for the anchor update. Reverse KL is mode-seeking, keeping the student close to confident text-rendering behavior and reducing drift during debiasing.

3.4 Training Objective

The objective combines both distillation losses with gold-response cross-entropy:

$$\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_{\text{anc}} \mathcal{L}_{\text{anc}} + \lambda_{\text{gold}} \mathcal{L}_{\text{gold}} \quad (9)$$

The teacher is frozen during cache construction, and the student is fine-tuned with LoRA (Hu et al., 2022). Training and hyperparameters are reported in Section 4.1 and Appendix A.1.

Model	Method	ConflictBank			MuSiQue			HotpotQA		
		Bias↓	DCR↑	Var.↓	Bias↓	DCR↑	Var.↓	Bias↓	DCR↑	Var.↓
Qwen3-8B	Base Model	0.276	0.309	0.0847	0.198	0.041	0.0014	0.127	0.045	0.0005
	Attn. Rebal. (Liu et al., 2026)	0.286	0.218	0.0643	0.156	0.050	0.0030	0.181	0.052	0.0008
	Vanilla KD	<u>0.214</u>	0.346	<u>0.0531</u>	0.179	0.058	0.0016	0.208	0.051	0.0001
	On-Policy KD	0.256	<u>0.408</u>	0.0744	<u>0.150</u>	<u>0.063</u>	0.0013	0.146	0.057	<u>0.0002</u>
	Multi-Fmt KD	0.264	0.319	0.0786	0.155	0.041	<u>0.0009</u>	<u>0.115</u>	0.034	<u>0.0002</u>
	FaDe (Ours)	0.210	0.420	0.0509	0.094	0.072	0.0008	0.093	<u>0.053</u>	0.0004
LLaMA-3.1-8B	Base Model	0.103	0.331	0.0140	0.077	0.053	0.0020	<u>0.220</u>	0.052	0.0008
	Attn. Rebal. (Liu et al., 2026)	0.091	0.398	0.0066	0.124	0.067	<u>0.0006</u>	0.230	0.060	0.0008
	Vanilla KD	0.094	0.459	<u>0.0058</u>	0.127	<u>0.072</u>	<u>0.0006</u>	0.291	0.067	0.0005
	On-Policy KD	<u>0.080</u>	<u>0.494</u>	0.0086	0.114	0.066	0.0012	0.267	<u>0.068</u>	<u>0.0003</u>
	Multi-Fmt KD	0.102	0.418	0.0153	<u>0.073</u>	0.051	0.0011	0.243	0.050	0.0008
	FaDe (Ours)	0.057	0.503	0.0048	0.068	0.074	0.0004	0.196	0.073	0.0002
Ministral-8B	Base Model	0.254	0.306	0.0742	0.224	0.265	0.0009	0.022	0.320	0.0004
	Attn. Rebal. (Liu et al., 2026)	0.090	0.284	<u>0.0030</u>	0.312	0.219	0.0009	0.106	0.253	0.0007
	Vanilla KD	<u>0.052</u>	0.366	0.0057	<u>0.221</u>	0.294	<u>0.0007</u>	0.023	0.356	0.0009
	On-Policy KD	0.088	<u>0.439</u>	0.0077	0.257	<u>0.307</u>	0.0015	0.035	0.367	0.0013
	Multi-Fmt KD	0.293	0.385	0.0923	0.239	0.292	0.0009	<u>0.020</u>	0.345	<u>0.0005</u>
	FaDe (Ours)	0.045	0.458	0.0007	0.205	0.346	0.0003	0.013	<u>0.358</u>	<u>0.0005</u>
Gemma-2-9B	Base Model	0.131	0.304	0.0236	0.046	0.095	<u>0.0019</u>	<u>0.191</u>	0.134	0.0007
	Attn. Rebal. (Liu et al., 2026)	0.117	0.375	0.0123	0.079	0.125	0.0020	0.246	0.132	0.0007
	Vanilla KD	0.099	0.469	<u>0.0052</u>	0.049	<u>0.132</u>	0.0020	0.238	<u>0.143</u>	0.0009
	On-Policy KD	<u>0.071</u>	<u>0.486</u>	0.0063	<u>0.040</u>	0.121	0.0027	0.228	0.142	<u>0.0006</u>
	Multi-Fmt KD	0.146	0.339	0.0257	0.047	0.070	0.0024	0.261	0.073	0.0008
	FaDe (Ours)	0.050	0.545	0.0035	0.029	0.163	0.0018	0.176	0.165	0.0005

Table 1: Cross-backbone and cross-dataset results. Bias is $|\text{FPR} - 0.5|$ ($|\Delta\text{FPR}|$). Lower Bias/FPR Var. is better; higher DCR indicates greater dual-source coverage. Best results are **bold**; second-best are underlined.

4 Experiments

We evaluate FaDe on three knowledge-conflict benchmarks and four backbones, then analyze component effects and counterfactual stability.

4.1 Experimental Setup

Datasets. We evaluate on ConflictBank (Su et al., 2024), MuSiQue (Trivedi et al., 2022), and HotpotQA (Yang et al., 2018). ConflictBank is filtered to 9,000 train and 3,000 test conflict pairs; MuSiQue and HotpotQA are converted by pairing supporting evidence with model-specific counter-evidence. Evidence is rendered as text, table, KG, or infobox, yielding six heterogeneous pairs. Appendix A.2 gives examples and construction audits.

Models. We use Qwen3-8B (Yang et al., 2025) as the primary backbone for main results and ablations. To test whether the trends are tied to a single model family, we also evaluate LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024), Ministral-8B-Instruct, and Gemma-2-9B-IT (Team et al., 2024), which are instruction-tuned open-weight models

of comparable scale but different training recipes. For each backbone, teacher and student start from the same checkpoint and the teacher is frozen, so differences reflect the distillation objective rather than extra model capacity.

Evaluation Metrics. Following Liu et al. (2026), we report Format Preference Rate (FPR)-based directional bias $|\text{FPR} - 0.5|$, Dual Coverage Rate (DCR), and FPR variance across format pairs. Lower is better for bias and variance; higher DCR indicates greater dual-source coverage. DCR measures whether a response acknowledges both conflicting sources, not factual accuracy. We interpret DCR jointly with directional bias, cross-format variance, and clean-QA retention rather than treating higher DCR alone as better final-answer correctness. Full-scale conflict results use deterministic rule-based adjudication for a uniform protocol. Appendix B.5 re-adjudicates a sampled subset with GPT-4o-mini majority voting and reports rule-LLM label transitions; this validation is diagnostic, not a replacement for the full protocol.

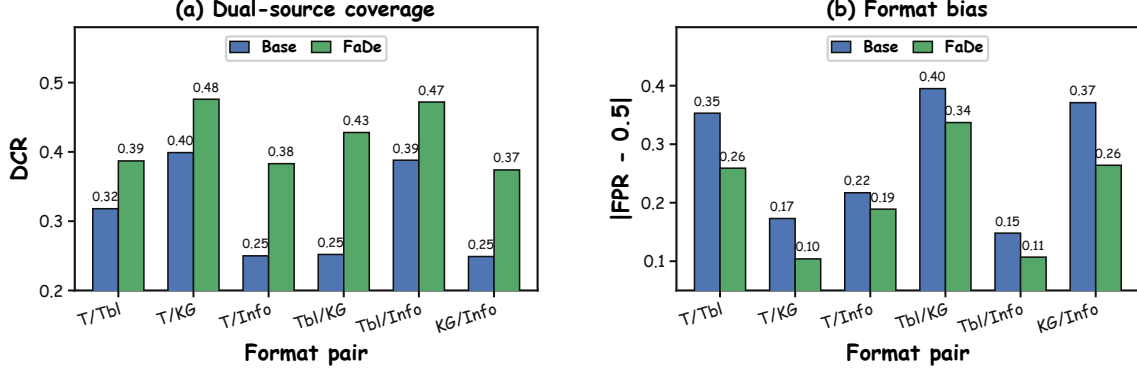


Figure 3: Bias-coverage trade-off between Base and FaDe on Qwen3-8B ConflictBank. Panels show per-pair DCR and $|FPR - 0.5|$; higher DCR means greater dual-source coverage, and lower bias means weaker format preference.

Method	Text / Table		Text / KG		Text / Infobox		Table / KG		Table / Infobox		KG / Infobox		Average	
	Bias↓	DCR↑	Bias↓	DCR↑	Bias↓	DCR↑	Bias↓	DCR↑	Bias↓	DCR↑	Bias↓	DCR↑	Bias↓	DCR↑
Base Model	0.353	0.318	0.173	0.399	<u>0.217</u>	0.250	0.395	0.252	0.148	0.388	0.371	0.249	0.276	0.309
On-Policy KD	<u>0.324</u>	<u>0.373</u>	<u>0.113</u>	<u>0.464</u>	0.223	<u>0.364</u>	<u>0.389</u>	<u>0.418</u>	0.140	<u>0.461</u>	<u>0.345</u>	<u>0.368</u>	<u>0.256</u>	<u>0.408</u>
Multi-Format KD	0.334	0.333	0.129	0.428	0.256	0.259	0.393	0.267	<u>0.123</u>	0.379	0.350	0.245	0.264	0.319
FaDe (Ours)	0.259	0.387	0.104	0.476	0.189	0.383	0.337	0.428	0.107	0.472	0.264	0.374	0.210	0.420

Table 2: Per-pair results on ConflictBank with Qwen3-8B. Bias is $|FPR - 0.5|$ ($|\Delta FPR|$). Lower Bias is better; higher DCR indicates greater dual-source coverage. The best result is **bold**; second-best is underlined.

Baselines. We compare FaDe against three baseline categories: (1) **Untuned/inference-time methods**: the Base Model and the lightweight attention-rebalancing intervention evaluated by Liu et al. (2026) (*Attn. Rebal.* in the tables). Both require no additional training. (2) **Distillation controls**: Vanilla KD (Offline) with fixed teacher soft targets and On-Policy KD on student-visited prefixes, implemented with standard KD objectives (Hinton et al., 2015; Agarwal et al., 2024). These reflect common distillation practice. (3) **Format-exposure control**: Multi-Format KD, which trains on all ordered heterogeneous assignments $\mathcal{F}_{\neq}^2$ but matches assignment-specific teacher predictions instead of marginalized targets. This clarifies the role of marginalization. Symmetric KL, Attention Balance, and Bidirectional Teacher Cache are mechanism-level controls reported in Appendix Table 28, as they are evaluated only in the focused Qwen3-8B ConflictBank setting.

Implementation Details. We train all students with LoRA (Hu et al., 2022) ($r=64$, $\alpha=128$), learning rate 3×10^{-6} , effective batch size 32, and 2 epochs. FaDe uses $M=7$ assignments and top- $K=256$ offline teacher caches with $\tau_{\text{low}}=0.01$, $\tau_{\text{high}}=0.05$. Full prompts, adjudication, cache construction, and compute details for reproduction are in Appendix A.1 and Appendix C.3.

4.2 Main Results

Table 1 reports cross-backbone and cross-dataset results for all methods, while Table 2 gives the per-format-pair breakdown on the main setting.

FaDe improves the bias-coverage trade-off. In the Qwen3-8B ConflictBank setting of Table 1, FaDe achieves the best trade-off. It reduces bias from 0.276 to 0.210 (-23.9%), raises DCR from 0.309 to 0.420, and lowers FPR variance from 0.0847 to 0.0509 (-39.9%). This joint improvement matters: a method could reduce directional bias by becoming uncertain or by overproducing neutral responses, but FaDe also increases the rate of responses that acknowledge both conflicting sources. Vanilla KD reduces bias but gives a smaller DCR gain, while On-Policy KD improves DCR but leaves higher bias and variance. This suggests that ordinary distillation objectives capture part of the teacher’s conflict behavior but do not remove the format-specific component that appears in individual renderings.

Focused mechanism-level controls in Appendix Table 28 show the same pattern: attention rebalancing and simple output/cache modifications do not match FaDe’s joint improvement. This supports the view that the key ingredient is the multi-format marginalization target rather than a generic symmetry or attention adjustment.

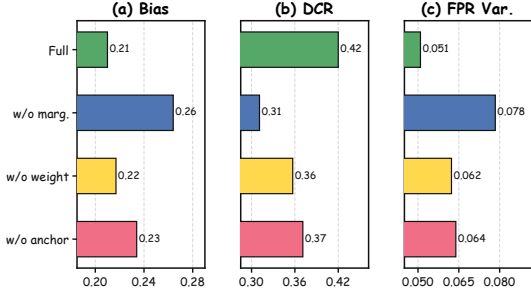


Figure 4: FaDe ablation on Qwen3-8B ConflictBank. Panels show Bias, DCR, and FPR variance; format marginalization is most critical.

FaDe concentrates improvement on the most biased pairs. Table 2 provides the per-format-pair breakdown behind Figure 3 and Figure 5. FaDe improves DCR on all six heterogeneous format pairs, showing that the aggregate gain is not driven by a single easy pair. The largest DCR gain appears on Table/KG (+0.176), where the base model has both high directional bias and low dual-source coverage. The largest relative bias reduction appears on Text/KG, from 0.173 to 0.104 (−39.9%), while KG/Infobox and Text/Table also show large absolute reductions. These patterns are consistent with the design of format marginalization: pairs with stronger teacher disagreement receive stronger adaptive debiasing, and the student learns a less format-dependent decision boundary. Multi-Format KD exposes the student to more ordered assignments, but because each assignment keeps its own teacher target, it can still preserve the teacher’s format preference rather than averaging it out. Figure 6 makes this mechanism explicit: format-pair bins with larger teacher JS disagreement receive stronger adaptive weights and show larger Base-to-FaDe gains than the low-disagreement bin.

FaDe is stable across backbones and controlled datasets. Table 1 evaluates whether the main finding holds beyond one backbone and one benchmark. Across 12 model–dataset settings, FaDe obtains the lowest bias in every setting and reduces bias by 34.1% on average. The largest reduction appears on Ministral-8B ConflictBank, where bias drops from 0.254 to 0.045 (−82.3%), but the trend is not limited to models with large initial bias: LLaMA-3.1 and Gemma also improve on ConflictBank, MuSiQue, and HotpotQA. The DCR improvements are especially clear on ConflictBank, where all four backbones gain substantially, and the cross-format variance consistently decreases. This matters because bias reduction alone can be achieved by collapsing to one safe response style;

Variant	$ \Delta\text{FPR} \downarrow$	DCR $\uparrow$	FPR Var. $\downarrow$
FaDe (full)	0.210	0.420	0.0509
w/o format margin.	0.264	0.311	0.0784
w/o adaptive weight.	0.217	0.357	0.0623
w/o anchoring path	0.234	0.371	0.0639

Table 3: Component ablation of FaDe on ConflictBank with Qwen3-8B. Each row removes one component from the full model to isolate its contribution.

the simultaneous DCR and variance gains suggest that FaDe changes format sensitivity while preserving more dual-source acknowledgement, although not necessarily final-answer correctness.

The QA-derived benchmarks provide a controlled stress test because their conflicts are constructed from originally clean multi-hop QA examples rather than from ConflictBank’s native conflict types. FaDe still achieves the lowest bias on MuSiQue and HotpotQA for every backbone, suggesting that the debiasing behavior also appears on controlled QA-derived conflict sets, while not establishing generalization to naturally occurring knowledge conflicts. Vanilla KD is competitive in selected cells but is less reliable: it increases HotpotQA bias for all four backbones and does not consistently improve DCR. Clean QA checks in Appendix C.2 show preserved Qwen3-8B F1 but a measurable drop for LLaMA-3.1, which we treat as an important capability-retention trade-off rather than hiding it in the main results.

A multi-pass format-ensembling diagnostic in Appendix C.1 provides a strong but costlier inference-time upper bound when repeated generations are allowed. FaDe targets a different deployment regime: it pays offline cache construction and LoRA fine-tuning once, then uses a reusable single-pass student at inference time.

4.3 Ablation Study

We use Appendix Table 28 and component removals in Table 3 to separate generic regularization from FaDe’s finite-support target construction.

Appendix controls show simpler replacements do not match FaDe. Table 3 and Figure 4 isolate FaDe’s components. Removing format marginalization causes the largest drop, increasing bias from 0.210 to 0.264 and reducing DCR from 0.420 to 0.311, supporting format averaging’s central role. Removing adaptive weighting lowers DCR to 0.357 and increases variance, indicating that disagreement-aware mixing helps preserve useful teacher signals while debiasing difficult samples.

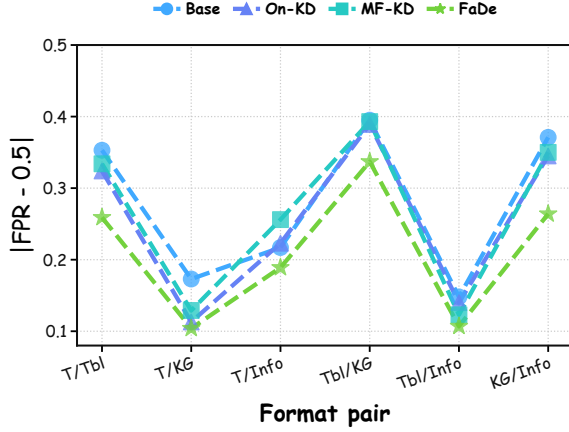


Figure 5: FaDe and baseline comparison across format pairs on Qwen3-8B ConflictBank. Lower $|FPR - 0.5|$ indicates weaker format preference.

Removing the anchoring path also weakens all three metrics, showing that all-text anchoring is part of the stable debiasing recipe. The components are complementary rather than redundant. Format marginalization supplies the debiased target, adaptive weighting decides how strongly to trust it, and anchoring prevents drift from stable text-format behavior. This decomposition explains why FaDe improves bias, DCR, and variance simultaneously.

4.4 Analysis

Disagreement Predicts Debiasing Gains. Figure 6 shows that higher teacher JS disagreement leads to larger adaptive weights and DCR gains, while bias reduction remains positive across bins. This supports disagreement-aware marginalization: FaDe applies stronger correction where format interventions reveal greater teacher instability. It also explains why fixed marginalization is suboptimal: stable samples need little correction, while unstable ones benefit from stronger averaging.

Residual Bias Varies Across Conflict Types. Across ConflictBank types, semantic conflicts retain the strongest residual bias ($|FPR - 0.5| = 0.070$), ahead of misinformation (0.051) and temporal conflicts (0.029), while DCR stays within 0.377–0.404. Residual errors therefore concentrate in harder cases where semantic disambiguation leaves more room for format shortcuts.

Counterfactual Format Stability. FaDe increases off-diagonal agreement across heterogeneous renderings of the same conflict, indicating improved counterfactual stability rather than a gain on a single biased pair. Higher cross-format agree-

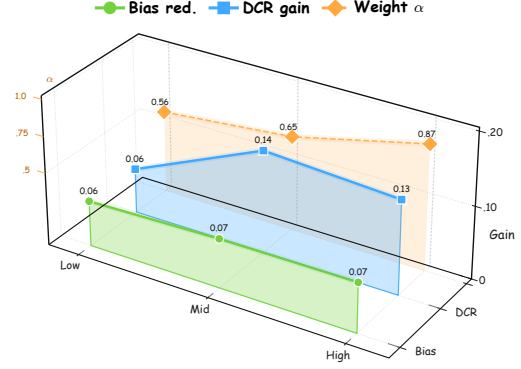


Figure 6: Debiasing gains by teacher disagreement. Higher-disagreement bins show stronger adaptive weights and larger DCR gains, with positive bias reduction.

Figure 6: Debiasing by teacher disagreement. Higher-disagreement bins show stronger adaptive weights and larger DCR gains, with positive bias reduction.

ment shows reduced sensitivity to the intervention itself, not just a shift toward another format. Appendix Figures 8 and 10 give details.

Format Effects Decrease Under Controls. After controlling for pair identity, order, and length, average absolute format coefficient drops from 0.226 to 0.162; format-following bias and flips decrease, while CFC rises from 0.459 to 0.570. These controls rule out simple order and length confounds for finite-support diagnostics, while not claiming unrestricted causal identifiability. Appendix Table 11 and Appendix B.4 give the full diagnostics.

5 Conclusion

We proposed FaDe, a cross-format framework that treats format as an assignable nuisance variable and constructs finite-support, format-marginalized teacher targets over controlled renderings. Its activation-and-anchoring design debiases heterogeneous inputs while preserving text behavior, yielding a reusable single-pass student after offline caching. Across three benchmarks and four backbones, FaDe lowers format bias; on Qwen3-8B ConflictBank, it reduces bias by 23.9%, lowers FPR variance by 39.9%, and raises DCR from 0.309 to 0.420. Ablations attribute the gain to format marginalization rather than format exposure, attention balancing, or symmetric-output constraints, and diagnostics show reduced sensitivity across controlled format assignments. These results support format marginalization as a route to reducing format sensitivity under predefined, semantically matched renderings. Future work should extend beyond predefined formats and binary conflicts.

Limitations

FaDe is evaluated on four formats and binary conflicts; code, charts, mixed documents, and multi-source disagreements remain open. It assumes formats can be enumerated before cache construction, so new deployment formats require extending the intervention support and refreshing the teacher cache. MuSiQue and HotpotQA conflicts are converted with model-specific, often synthetic counter-evidence; they should be viewed as controlled stress tests for format sensitivity rather than evidence of naturally occurring conflict generalization. Human semantic audits would strengthen this evaluation. Automatic adjudication may mislabel borderline answers; Appendix B.5 reports an LLM re-adjudication diagnostic, but it is still not a substitute for a large human-labeled audit. Capability retention is empirical and targeted: Appendix C.2 shows preserved Qwen3-8B F1 but lower LLaMA-3.1 F1 on clean QA, suggesting that retention may require backbone-specific tuning. FaDe adds offline cache cost, and broader instruction following, reasoning, and safety capabilities remain to be checked. Finally, FaDe reduces but does not eliminate format bias; Appendix B.5 shows out-of-support gains should not be assumed.

Ethics Statement

All datasets are public benchmarks cited in the paper, and we do not collect new user data. QA-derived counter-evidence is used only for robustness evaluation and should not be treated as factual or redistributed as verified knowledge. We use open-weight LLMs and LoRA fine-tuning under relevant licenses. FaDe reduces format-driven preference, but does not guarantee factual correctness, eliminate hallucinations, or suit high-stakes domains without human verification. A higher tendency to acknowledge both conflicting sources should not be interpreted as resolving which source is true. Users should ensure trustworthy evidence, especially in medicine, finance, law, and public policy. Format debiasing may leave source-credibility, language, cultural, or retrieval biases unaddressed.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful comments and helpful suggestions, which helped improve this work.

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant

No. 62576256).

References

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.
- Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.
- Lekai Chen, Ashutosh Trivedi, and Alvaro Velasquez. 2024. LLMs as probabilistic minimally adequate teachers for DFA learning. *arXiv preprint arXiv:2408.02999*.
- Lekai Chen, Ashutosh Trivedi, and Alvaro Velasquez. 2026. Towards persistent noise-tolerant active learning of regular languages with class query. In *International Conference on Learning Representations*, volume 2026, pages 137832–137855.
- Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models’ local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.
- 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 models. *arXiv preprint arXiv:2407.21783*.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pages 32694–32717.
- Yu Guo, Shenghao Ye, Shuangwu Chen, Zijian Wen, Tao Zhang, Bai Qirui, Dong Jin, Yunpeng Hou, Huasen He, Jianyang, and Xiaobin Tan. 2026. Rethinking table pruning in TableQA: From sequential revisions to gold trajectory-supervised parallel search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12960–12976.

- Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4320–4333.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. WikiContradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from Wikipedia. In *Advances in Neural Information Processing Systems*, volume 37, pages 109701–109747.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2021–2031.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Jiacheng Liu, Mayi Xu, Qiankun Pi, Wenli Li, Ming Zhong, Yuanyuan Zhu, Mengchi Liu, and Tiejun Qian. 2026. Format as a prior: Quantifying and analyzing bias in LLMs for heterogeneous data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32060–32068.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.
- Siya Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025. Open domain question answering with conflicting contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1838–1854.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063.
- Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695.
- Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.
- Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2022. UniK-QA: Unified representations of structured and unstructured knowledge for open-domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1535–1546.
- Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4902–4912.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 103242–103268.
- Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4231–4242.
- Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

- Wenxin Tang, Jingyu Xiao, Yanpei Gong, Fengyuan Ran, Tongchuan Xia, Junliang Liu, Man Ho Lam, Wenxuan Wang, and Michael R. Lyu. 2026a. EfficientPosterGen: Semantic-aware efficient poster generation via token compression and accurate violation detection. *arXiv preprint arXiv:2603.00155*.
- Wenxin Tang, Jingyu Xiao, Zhenyu Liu, Zipeng Xie, Junliang Liu, Wang Luo, Yuan Jiang, Yintong Huo, and Michael Lyu. 2026b. CodeShrink: Adaptive visual compression for efficient multimodal code understanding. *arXiv preprint arXiv:2607.29637*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610.
- Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.
- Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. F-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834.
- Zhengxian Wu, Hangrui Xu, Kai Shi, Zhuohong Chen, Yunyao Yu, Chuanrui Zhang, Zirui Liao, Jun Yang, Zhenyu Yang, Haonan Lu, and Haoqian Wang. 2026. ProMSA: Progressive multimodal search agents for knowledge-based visual question answering. *arXiv preprint arXiv:2606.27974*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, volume 2024, pages 35623–35646.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.
- Shenghao Ye, Yu Guo, Dong Jin, Yuxiang Wang, Yikai Shen, Yunpeng Hou, Shuangwu Chen, Jianyang, and Xiaofeng Jiang. 2026a. When TableQA meets noise: A dual denoising framework for complex questions and large-scale tables. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24022–24045.
- Shenghao Ye, Yuxiang Wang, Yu Guo, Dong Jin, Shuangwu Chen, and Jian Yang. 2026b. Rethinking stepwise model routing: A cost-efficient table reasoning perspective. *arXiv preprint arXiv:2605.29319*.
- Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8413–8426.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in neural information processing systems*, 36:46595–46623.

A Reproducibility and Data Construction

This appendix is organized into three parts. We first provide reproducibility details and data-construction checks, then report robustness and finite-support intervention diagnostics, and finally collect additional controls, capability checks, and computational overhead measurements. Appendix tables use a consistent single-column booktabs style where possible: compact spacing, fixed-width text columns, and bold FaDe rows without green shading.

Appendix road map. The appendix is meant to be read as a verification path rather than a result dump: implementation settings come first, dataset construction checks follow, and the later sections isolate robustness, finite-support intervention, stress-test, capability, and cost questions.

(a) Training and cache settings	
Hyperparameter	Value
LoRA rank r	64
LoRA scaling α	128
Learning rate	3×10^{-6}
Batch size	2
Gradient accumulation steps	16
Effective batch size	32
Training epochs	2
Top- K (teacher cache)	256
Format assignments M	7: all-text control plus six heterogeneous pairs
τ_{low}	0.01
τ_{high}	0.05
α_{min}	0.0
α_{max}	1.0
λ_{anc}	1.0
λ_{gold}	0.2

(b) Conflict-resolution prompt skeleton	
Field	Template line
Question	Question: {question}
Evidence 1	{evidence_A} rendered as f_A
Evidence 2	{evidence_B} rendered as f_B
Instruction	Based on the above evidence, provide your answer.

Table 4: Reproducibility setup for FaDe. Panel (a) lists the training and cache-construction settings; panel (b) gives the conflict-resolution prompt skeleton used for both distillation and evaluation.

A.1 Training, Prompting, and Cache Construction

Pipeline guide. This subsection follows the actual execution order: scalar training settings and the prompt skeleton, offline teacher-cache construction, and then generation-time adjudication. The table fixes values; the algorithms define the operations that consume them.

The reproducibility material is arranged in the same order as the implementation pipeline: Table 4 first fixes the scalar training settings and the exact prompt skeleton; Algorithm 1 then describes how the teacher cache is built; Algorithm 2 finally specifies how generated responses are adjudicated into the metrics used throughout the paper. This layout is intended to make the appendix executable rather than merely descriptive.

Algorithm 1 summarizes the offline target-construction step. We spell out the procedural component separately from Table 4 so that the reader can distinguish cached teacher quantities, per-format interventions, and the adaptive mixture used by student training. The teacher is frozen throughout this step; all expensive teacher-forced distributions are computed once, compressed into top- K caches, and reused during student training.

The algorithm complements the scalar setup: the table fixes the values, whereas the pseudocode clarifies where Eq. 4 and the adaptive mixture in Eq. 6

Algorithm 1: Format-Marginalized Teacher Target Construction

```

Input      :  $c_i, f_i^{\text{orig}}, \mathcal{A}_{\text{eval}}$ , frozen  $T$ , prefix  $y_i$ , top- $K$ , thresholds
Output    : Top- $K$  target cache  $\mathcal{C}_i$  and adaptive weight  $\alpha_i$ 

 $\mathcal{A}_{\text{het}} \leftarrow \mathcal{A}_{\text{eval}} \setminus \{( \text{text}, \text{text} )\}$ 
 $P_T^{\text{text}} \leftarrow T(\text{render}(c_i, \text{text}, \text{text}), y_i)$ 
 $P_T^{\text{orig}} \leftarrow T(\text{render}(c_i, f_i^{\text{orig}}), y_i)$ 

Initialize log-average log  $\bar{P}_T$  and disagreement list  $\mathcal{D}$ 
foreach  $a_j = (f_A^{(j)}, f_B^{(j)}) \in \mathcal{A}_{\text{eval}}$  do
     $x_j \leftarrow \text{render}(c_i, a_j)$ 
     $P_T^{(j)} \leftarrow T(x_j, y_i)$ 
    Update log  $\bar{P}_T$  with Eq. 4
    if  $a_j \in \mathcal{A}_{\text{het}}$  then
        Append  $D_{\text{JS}}(P_T^{\text{text}} \| P_T^{(j)})$  to  $\mathcal{D}$ 
    end
end

 $d_i \leftarrow \text{mean}(\mathcal{D})$ 
 $\alpha_i \leftarrow \text{clip}((d_i - \tau_{\text{low}}) / (\tau_{\text{high}} - \tau_{\text{low}}), \alpha_{\text{min}}, \alpha_{\text{max}})$ 
 $\bar{P}_{T,i} \leftarrow \alpha_i \bar{P}_{T,i} + (1 - \alpha_i) P_{T,i}^{\text{orig}}$ 
 $\mathcal{C}_i \leftarrow \text{TopK}(\bar{P}_{T,i}, K)$ , then renormalize
Return  $\mathcal{C}_i, \alpha_i$ 

```

Algorithm 2: Conflict Evaluation and Adjudication Protocol

```

Input      : Pairs  $\mathcal{D}$ , format set  $\mathcal{P}$ , model  $M$ , seed  $s$ , adjudicator  $A$ 
Output    : Bias, DCR, and FPR variance

Initialize response records  $\mathcal{R} \leftarrow \emptyset$ 
foreach semantic conflict pair  $q_i \in \mathcal{D}$  do
    foreach format pair  $(f, g) \in \mathcal{P}$  do
        Sample evidence order using seed  $s$ 
         $x_i^{(f,g)} \leftarrow \text{renderPrompt}(q_i, f, g)$ 
         $\hat{y}_i^{(f,g)} \leftarrow M(x_i^{(f,g)})$ 
         $\ell_i^{(f,g)} \leftarrow A(\hat{y}_i^{(f,g)}) \in \{A, B, \text{Both}, \text{Neither}\}$ 
        Add  $(q_i, f, g, \ell_i^{(f,g)})$  to  $\mathcal{R}$ 
    end
end

Compute FPR for each format pair from single-source labels in  $\mathcal{R}$ 
Compute DCR from labels equal to Both
Compute FPR variance across format pairs
Return  $|\Delta \text{FPR}|, \text{DCR}, \text{FPR variance}$ 

```

enter the cache. Student training then consumes $\mathcal{C}_i$ in the activation path and uses the original/text path as the anchor described in Section 3.3.

For evaluation, each knowledge conflict sample is rendered with the same prompt skeleton in Table 4, but the operation changes from teacher-forced cache construction to free-form generation followed by response classification. Algorithm 2 makes this distinction explicit. This is important because the bias and DCR metrics are computed over adjudicated responses, not over teacher probabilities.

A.2 Conflict Dataset Examples and Audit

Data checks. The example table and audit table answer different questions. The examples make the semantic conflicts concrete, while the audit checks whether the rendered files contain the answer-bearing strings needed for rule-based evaluation.

Table 5 shows representative claim pairs after

Dataset	Conflict summary
ConflictBank	Q: What position does Guilhem Cledele currently or formerly hold? A: conseiller municipal de Montvalent; B: cardinal in 2043. Type: temporal.
MuSiQue	Q: What is the capital of province containing Hartashen? A: Kapan; B: Hassankeyf. Type: counter-answer.
HotpotQA	Q: What instrument did Nigel Walley play in John Lennon’s band? A: tea-chest bass; B: bass guitar. Type: counter-answer.

Table 5: Representative conflict-construction examples. Lists paired claims; evidence passages are rendered in text, table, KG, and infobox formats during evaluation.

conversion. For QA-derived datasets, the conflicting side is constructed from model-specific counter-answers and then rendered into the same predefined format set. The other backbone-specific evaluation files follow the same construction procedure with the corresponding local filter model. The table renderings here are compact passage-level tables of a few hundred tokens on average (Table 6); reasoning over large or noisy tables is a separate challenge addressed by table denoising and pruning methods (Ye et al., 2026a; Guo et al., 2026), which FaDe does not target.

Table 6 adds automatic validity checks for the converted conflict sets used in the primary Qwen3-8B evaluation. The checks verify that required fields and rendered formats are present, that each side’s answer string is supported by its own text evidence, and that answer strings do not create obvious substring-containment risks for rule adjudication. The QA-derived datasets have high completion and answer-support rates; their counter-evidence is often synthetic by construction because it is built from model-specific wrong answers. The final columns report average rendered-token lengths, which are not normalized across formats. This length variation is a remaining evaluation confound, but it is shared by all compared methods because they are evaluated on the same rendered prompts.

Because the automatic audit cannot judge semantic compatibility, Table 7 adds a small model-assisted audit of the QA-derived conflicts. We sample 50 MuSiQue and 50 HotpotQA conflicts and ask LLaMA-3.1-8B-Instruct to check four conditions: both sides address the same question, the supporting side entails Answer A, the counter-evidence side states or supports Answer B, and the two answers are incompatible. The audit shows

Dataset	Pairs	Quality checks	Avg. tokens
ConflictBank	3000	Complete 100.0; supported 97.0; synthetic 0.0; containment risk 0.1.	Text/Table/KG/Info. = 675/157/306/90; max/min 7.52.
MuSiQue	3000	Complete 100.0; supported 100.0; synthetic 100.0; containment risk 0.0.	Text/Table/KG/Info. = 239/311/269/277; max/min 1.31.
HotpotQA	3000	Complete 100.0; supported 98.9; synthetic 79.8; containment risk 0.0.	Text/Table/KG/Info. = 191/258/219/226; max/min 1.35.

Table 6: Automatic audit of Qwen3-8B conflict-evaluation data. Complete = required claims, answers, text evidence, and structured renderings are non-empty. Supported = answer A appears in evidence A and answer B in evidence B under normalized matching. Contain. risk = one normalized answer string contains the other. Avg. tokens are Text / Table / KG / Infobox.

Dataset	N	Same Q	Ans. conf.	Sup. A	Ctr. B	Strict	Synth.
MuSiQue	50	100.0	89.8	91.8	71.4	53.1	100.0
HotpotQA	50	100.0	94.0	100.0	54.0	48.0	72.0
All	100	100.0	91.9	96.0	62.6	50.5	86.0

Table 7: Model-assisted audit of QA-derived conflict construction. Values are percentages over sampled Qwen3 evaluation conflicts. Same Q checks whether both sides address the same question; Ans. conflict checks whether the two answers are incompatible; Support A and Counter B check whether the corresponding evidence side supports or states its answer. Strict all requires all four conditions and is intentionally conservative. The audit supports using MuSiQue and HotpotQA as controlled stress tests, while confirming that synthetic counter-evidence limits natural-conflict generalization.

strong question alignment and answer incompatibility, while the stricter all-condition score is lower because the counter-evidence side is often a synthetic claim statement rather than a naturally retrieved passage. This supports using the QA-derived sets as controlled stress tests for format sensitivity while reinforcing that they should not be read as naturally occurring conflict benchmarks.

Taken together, the construction examples and audit tables serve different purposes. Table 5 makes the semantic conflict construction concrete, Table 6 checks that the rendered evaluation files are complete enough for the format-bias measurements used in the main paper, and Table 7 probes whether the QA-derived counter-evidence is semantically usable. These audits are also useful for interpreting later stress tests: the QA-derived datasets are intentionally synthetic on one side, so their role is to test whether a model follows contradictory evidence under controlled rendering rather than to estimate naturally occurring factual disagreement.

Method	$ \Delta\text{FPR} \downarrow$	DCR $\uparrow$	FPR Var. $\downarrow$
Base Model	0.276 [0.263, 0.287]	0.309 [0.292, 0.328]	0.0847 [0.0783, 0.0911]
Vanilla KD (Offline)	0.214 [0.208, 0.228]	0.346 [0.329, 0.361]	0.0531 [0.0481, 0.0592]
Attn. Rebal. (Liu et al., 2026)	0.286 [0.271, 0.295]	0.218 [0.201, 0.236]	0.0643 [0.0609, 0.0724]
Multi-Format KD	0.264 [0.252, 0.277]	0.319 [0.300, 0.337]	0.0786 [0.0723, 0.0854]
FaDe (Ours)	0.210 [0.194, 0.223]	0.420 [0.399, 0.443]	0.0509 [0.0446, 0.0576]

Table 8: Bootstrap 95% confidence intervals on Qwen3-8B ConflictBank. Intervals are computed over 1,000 bootstrap resamples of semantic pair ids while keeping all rendered format variants for each sampled pair.

B Robustness, Finite-Support Diagnostics, and Stress Tests

Diagnostic map. Each block targets one threat to validity: bootstrap intervals test sampling uncertainty, order seeds test position confounding, CFC and transition analysis test counterfactual stability, fixed-effect and support checks test whether format dependence decreases under controlled assignments, and stress tests mark the boundary of the predefined format support.

B.1 Additional Result Breakdowns

Table 2 in the main text provides the detailed per-format-pair results behind the main bias–coverage figures.

Figure 7 visualizes the cross-backbone and cross-dataset results summarized in Table 1. The heatmaps show that FaDe’s largest relative bias reductions occur on ConflictBank, while the DCR improvements remain positive across datasets.

Figure 8 visualizes the counterfactual agreement matrix discussed in Section 4.4. FaDe increases agreement across all off-diagonal format-pair comparisons, indicating more stable decisions under format interventions.

B.2 Uncertainty and Order Robustness

Table 8 reports 95% bootstrap confidence intervals by resampling semantic conflict pairs and keeping all rendered format variants for each sampled pair. The intervals support the main-table trend that FaDe reduces both average format bias and cross-format variance relative to the base model and Multi-Format KD while improving DCR.

The main evaluation uses seeded random A/B order swaps to avoid conflating format preference with position preference. Table 9 reruns the Qwen3-8B ConflictBank evaluation on a 300-pair diagnostic subset with three order-randomization seeds. FaDe remains consistently less format-biased than

Method	$ \Delta\text{FPR} \downarrow$	DCR $\uparrow$	Var. $\downarrow$
Base Model	0.287 ± 0.004	0.323 ± 0.005	0.0919 ± 0.0012
FaDe (Ours)	0.215 ± 0.004	0.430 ± 0.001	0.0545 ± 0.0003

Table 9: Order-randomization seed robustness on a 300-pair Qwen3-8B ConflictBank diagnostic subset. Means and standard deviations over seeds 7, 13, and 42.

the base model and has small seed-level variance, suggesting that the main conclusion is not driven by a particular random ordering.

Figure 9 gives a more direct view of the same phenomenon. Each off-diagonal cell measures how often the row format is preferred against the column format after excluding dual-source responses. Values near 0.5 indicate weak directional preference; values close to 0 or 1 indicate strong format dominance. FaDe pulls most off-diagonal cells toward 0.5, especially for text/table and KG/infobox comparisons, showing that the aggregate bias reduction in Tables 8 and 9 is not caused by a single isolated format pair.

Together, Tables 8 and 9, together with Figure 9, show that the reported gains are stable under two forms of resampling and visible at the pairwise format level. The bootstrap analysis varies which semantic conflicts are included, while the order-seed diagnostic varies the A/B presentation order. FaDe remains lower-bias in both views, which reduces the risk that the main result is an artifact of a particular sample or a fixed passage order.

B.3 Counterfactual Stability and Transitions

To directly test whether predictions are stable under format interventions, we compute Counterfactual Format Consistency (CFC): for each semantic conflict pair, we compare the response classifications over the six heterogeneous format renderings and report the average pairwise agreement. We also report the all-format agreement rate and the rate at which a sample flips between preferring Claim A and preferring Claim B across formats. Table 10 shows that FaDe improves CFC over the base model and Multi-Format KD, and reduces single-side flips relative to the base model. Vanilla KD has the highest CFC in this diagnostic, but its dual-label rate remains below FaDe and its cross-dataset bias behavior is worse in the main evaluation.

Figure 10 decomposes how predictions change from the base model to FaDe for the dominant single-source decision in each heterogeneous for-

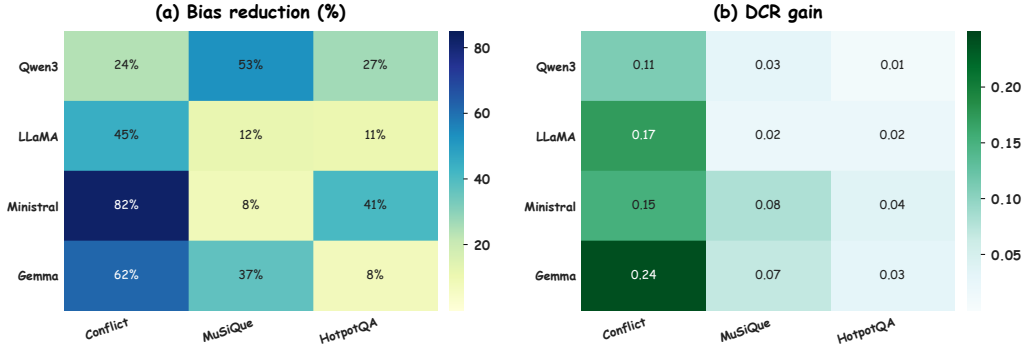


Figure 7: Cross-backbone and cross-dataset results. Panel (a) reports relative reduction in $|\Delta\text{FPR}|$ from the base model to FaDe; panel (b) reports the corresponding DCR gain. Values are computed from Table 1.

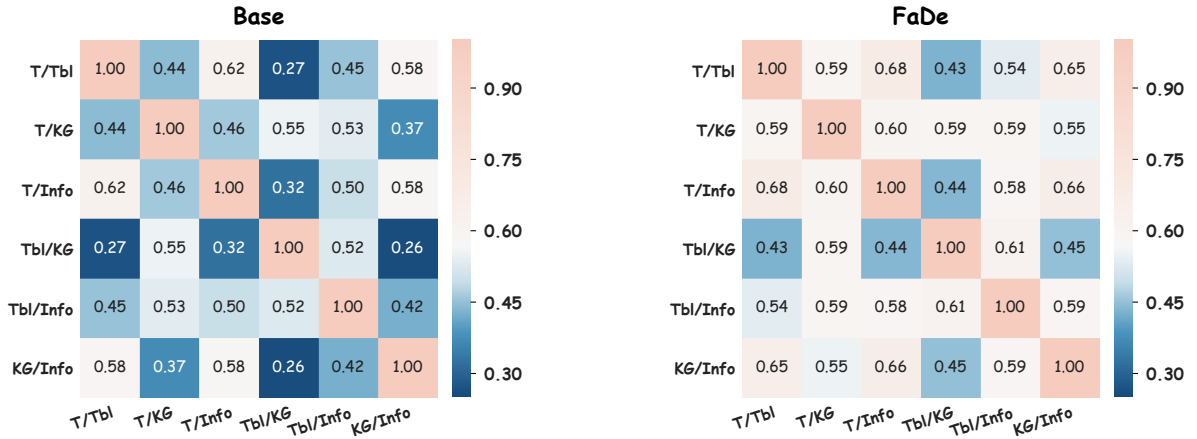


Figure 8: Counterfactual agreement between format interventions on Qwen3-8B ConflictBank. Each cell measures how often the response classification for the same semantic conflict agrees under two heterogeneous format pairs. FaDe increases agreement across all off-diagonal comparisons.

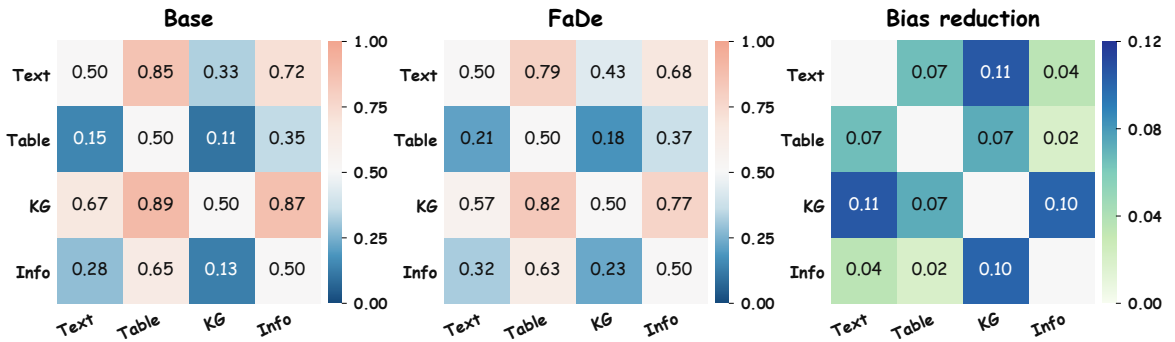


Figure 9: Directed format-preference matrices on Qwen3-8B ConflictBank. Off-diagonal cells give the probability that the row format is preferred over the column format. The right panel shows distance-to-0.5 reduction after FaDe.

mat pair. For each pair, we first identify which side the base model prefers more often, then track whether those originally dominant decisions remain dominant, flip to the opposite side, move to the dual-source label, or become neither. FaDe’s bias reduction comes from both direct rebalancing and increased dual-source recognition. The strongest shifts to dual-source responses occur on

table / KG and KG / infobox, which are also among the most biased base-model pairs.

Table 10 and Figure 10 should be read together. The CFC metric summarizes whether labels are stable across counterfactual renderings of the same content, whereas the transition figure explains how FaDe changes the base model’s dominant single-source decisions. This distinction matters because

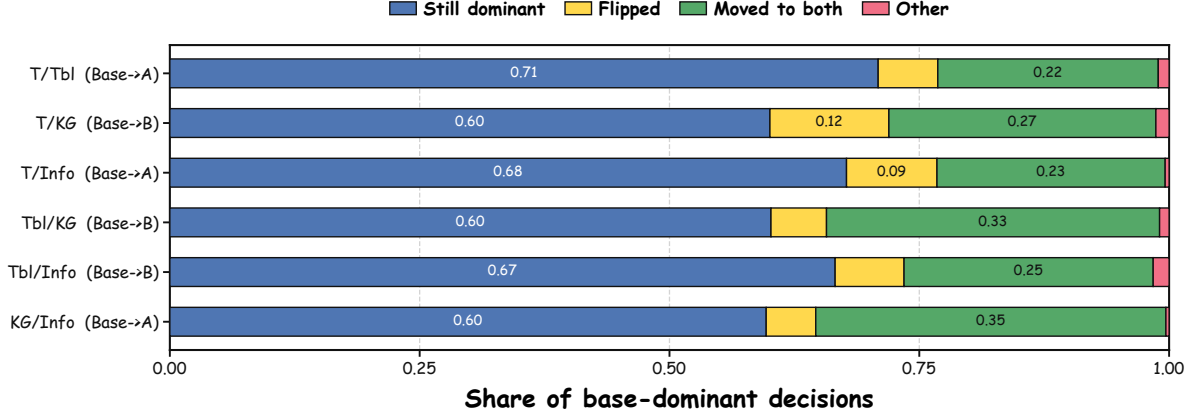


Figure 10: Transition decomposition from Base to FaDe on Qwen3-8B ConflictBank. Each stacked bar conditions on examples where the base model made the dominant single-source decision for that format pair. The green segment shows cases that move to dual-source coverage under FaDe.

Method	Pairs	CFC↑	All agree↑	A/B flip↓	Dual↑
Base Model	1000	0.459	0.069	0.605	0.309
Vanilla KD (Offline)	1000	0.620	0.291	0.293	0.346
Multi-Format KD	1000	0.482	0.088	0.573	0.319
FaDe (Ours)	1000	0.570	0.207	0.419	0.420

Table 10: Counterfactual format consistency on Qwen3-8B ConflictBank. CFC is pairwise classification agreement over six heterogeneous format renderings of the same conflict. Higher is better for CFC and all-format agreement; lower is better for single-side flips.

a method could look consistent by preserving a biased decision; FaDe instead increases consistency while also moving many cases toward dual-source recognition.

B.4 Finite-Support Intervention Diagnostics

Diagnostic logic. This subsection separates three possible confounds within the predefined intervention support: whether format coefficients persist after content, order, and length controls; whether a preferred format stays preferred after side swaps; and whether the finite intervention support is fully covered by the diagnostic renderings.

Table 11 summarizes the main finite-support intervention diagnostics used in Section 4.4.

Table 12 reports a content-fixed-effect linear probability model over single-source decisions. The dependent variable is whether the model prefers Claim A. The format coefficients are contrast-coded against text, and the regression controls for whether Claim A appears second and for rendered length difference. FaDe reduces the magnitude of the table and infobox coefficients and makes the KG coefficient indistinguishable from zero under the bootstrap interval.

Table 13 evaluates all 16 ordered assignments

Diagnostic	Base	FaDe	Change
Controlled format effect ↓	0.226	0.162	−28.3%
Format-following bias ↓	0.261	0.209	−19.8%
Paired single-side flip ↓	0.256	0.163	−36.1%
CFC agreement ↑	0.459	0.570	+0.111

Table 11: Finite-support intervention diagnostics on Qwen3-8B ConflictBank. Controlled format effect is the mean absolute coefficient of table/KG/infobox indicators in a content-fixed-effect linear probability model with order and length controls. Format-following bias is the mean absolute deviation from 0.5 when checking whether a fixed format remains preferred after swapping sides. Lower is better except for CFC. These diagnostics test reduced format dependence under controlled assignments, not unrestricted causal identifiability.

Variable	Base Model	FaDe (Ours)
Table vs. Text	-0.357 [-0.407,-0.303]	-0.269 [-0.339,-0.186]
KG vs. Text	+0.087 [+0.044,+0.131]	+0.038 [-0.021,+0.105]
Infobox vs. Text	-0.235 [-0.297,-0.169]	-0.179 [-0.265,-0.072]
Claim A second	-0.016 [-0.041,+0.011]	-0.166 [-0.199,-0.134]
Length diff. / 100 tok.	-0.030 [-0.047,-0.012]	-0.014 [-0.037,+0.017]

Table 12: Content-fixed-effect regression with order and length controls on Qwen3-8B ConflictBank. Values are coefficients with 95% bootstrap intervals. Format coefficients are contrast-coded against text; length difference is measured in hundreds of rendered tokens.

over a 500-pair diagnostic subset with fixed evidence order. For each unordered format pair, we compare (f, g) and (g, f) renderings of the same content and measure whether the same format remains preferred after swapping sides. FaDe reduces the dominant-format tracking rate for five of the six pairs and lowers paired single-side flips for all six pairs.

Table 14 stratifies heterogeneous renderings by the absolute log length ratio between the two evi-

Format pair	Dominant	Track↓ Base/FaDe	Flip↓ Base/FaDe
text/table	text	0.817 / 0.747	0.250 / 0.196
text/kg	kg	0.629 / 0.582	0.124 / 0.088
text/infobox	text	0.732 / 0.702	0.292 / 0.230
table/kg	kg	0.902 / 0.813	0.316 / 0.156
table/infobox	infobox	0.585 / 0.589	0.124 / 0.104
kg/infobox	kg	0.903 / 0.822	0.428 / 0.206
Average	–	0.761 / 0.709	0.256 / 0.163

Table 13: Format-following diagnostics on 500 Qwen3-8B ConflictBank pairs under all 16 ordered assignments. Tracking is how often the dominant format remains preferred across side swaps; flip is the rate at which single-source decisions switch sides between (f, g) and (g, f) . Lower is better for tracking and flip.

Length bin	N	Bias↓ Base/FaDe	DCR↑ Base/FaDe
Low	2001	0.258 / 0.176	0.413 / 0.523
Mid	2000	0.274 / 0.178	0.312 / 0.418
High	1999	0.259 / 0.236	0.280 / 0.389

Table 14: Length-stratified Qwen3-8B ConflictBank diagnostics. Bins use the absolute log length ratio between rendered evidence sides. Bias denotes $|\Delta\text{FPR}|$.

dence sides. FaDe reduces format bias in all three strata, although the improvement is smaller in the highest length-imbalance bin. This confirms that the main reduction is not limited to length-matched cases.

Table 15 removes the length signal more directly. Using the same per-sample adjudication records, we keep only pairs whose two rendered evidence sides are close in token length, under three matching thresholds on the absolute log length ratio (≤ 0.1 , 0.2 , and 0.3 , i.e., within roughly 10%, 22%, and 35%), and recompute the metrics on each subset. FaDe still reduces bias, raises DCR, and lowers FPR variance at every threshold, even though little length signal remains to exploit; the margin is smallest on the strictest and smallest subset. Together with the explicit length covariate in Table 12 and the stratified view in Table 14, this indicates that the measured debiasing reflects invariance to format structure rather than a density shortcut, while the position shortcut is excluded by order randomization (Table 9).

Table 16 combines two input-support checks that are easy to conflate. The rendering-validity row audits whether the predefined renderings are complete and preserve each side’s answer string across all formats. The all-16 diagnostic rows verify that the broader ordered diagnostic support is complete for each sampled content pair and report homogeneous-format controls. These rows are not

Length match	N	Bias↓ Base/FaDe	DCR↑ Base/FaDe	FPR Var.↓ Base/FaDe
$ \log \text{ratio} \leq 0.1$	302	0.250 / 0.242	0.404 / 0.490	0.0848 / 0.0768
$ \log \text{ratio} \leq 0.2$	604	0.231 / 0.221	0.422 / 0.477	0.0709 / 0.0596
$ \log \text{ratio} \leq 0.3$	901	0.244 / 0.180	0.413 / 0.478	0.0672 / 0.0335

Table 15: Length-matched subsets of Qwen3-8B ConflictBank. Each row keeps only pairs whose two rendered evidence sides fall within the given absolute log length ratio and recomputes the metrics on that subset. Bias denotes $|\Delta\text{FPR}|$.

Check	Setting	Coverage	Outcome
Rendering validity	All formats, 3000 pairs	100.0 complete	90.5 own support; 1.8 leakage
All-16 diagnostic controls	Base Model, 500 pairs	100.0 coverage	0.078 hom. bias; 0.572 hom. DCR
All-16 diagnostic controls	FaDe (Ours) , 500 pairs	100.0 coverage	0.174 hom. bias; 0.534 hom. DCR

Table 16: Rendering-support and all-16 diagnostic checks on Qwen3-8B ConflictBank. Own support requires each side’s answer string to appear in all four renderings of that side; leakage means at least one side contains the opposite answer string. Coverage is the fraction of diagnostic content pairs with all 16 ordered assignments present.

the FaDe target-construction support; they are a boundary sanity check confirming that same-format interventions are included in the diagnostic even though the method is designed for conflicts where two sources appear in different surface formats.

The key point is that rendering validity and support coverage address different failure modes. The first check asks whether the input construction preserves the answer-bearing strings needed for adjudication; the second asks whether the diagnostic spans the broader all-16 ordered support used for side-swap and homogeneous-format checks, not whether FaDe trains on all 16 assignments. FaDe’s weaker homogeneous-format row is therefore not interpreted as a contradiction to the main claim. It marks a setting where there is no cross-format conflict to marginalize.

Support-set and weighting sensitivity. The seven training assignments are the complete heterogeneous coverage of the four formats, namely one all-text control plus all $\binom{4}{2} = 6$ heterogeneous pairs, rather than an arbitrary subset. Homogeneous pairs contain no cross-format conflict to marginalize, and presentation order is handled by seeded A/B randomization at evaluation time (Table 9) and probed by the all-16 ordered diagnostics (Table 13), so including order in the support would more than double the cache cost. Table 17

Variant	M	Bias↓	DCR↑	FPR Var.↓
FaDe (Ours): default support	7	0.210	0.420	0.0509
Uniform averaging	7	0.217	0.357	0.0623
All 16 ordered assignments	16	0.206	0.442	0.0487

Table 17: Support-set and weighting variants on Qwen3-8B ConflictBank. M is the number of format assignments in the marginalization support; the default support is one all-text control plus the six heterogeneous pairs. The uniform-averaging row replaces the disagreement-aware weight with a uniform average and corresponds to the w/o adaptive weight ablation in Table 3.

tests whether these choices drive the outcome by re-training Qwen3-8B on ConflictBank with the same budget and protocol as the main setting, changing only the marginalization support or the weighting scheme. Uniform averaging loses much of the DCR gain, indicating that the disagreement-aware weight secures dual-source coverage, while enlarging the support to all 16 ordered assignments (at roughly $2.3\times$ cache cost) changes the outcome only marginally. The compact support is therefore not a sensitive choice.

B.5 Adjudicator and Out-of-Support Stress Tests

Stress-test scope. These results should be read as boundary checks. LLM re-adjudication tests whether the deterministic labels drive the conclusion, while the held-out JSON wrapper tests whether finite-support marginalization transfers to an unseen surface format.

This subsection connects two caveats from the main text: automatic adjudication is a scalable but imperfect proxy, and FaDe marginalizes over a predefined finite support rather than arbitrary future formats. We therefore use these tests to delimit the claims, not to introduce a stronger benchmark setting.

Table 18 groups the two stress tests that define the boundary of the current evaluation protocol. The LLM-adjudicator diagnostic re-adjudicates 300 sampled responses with GPT-4o-mini majority voting and asks whether the main conclusion depends on deterministic string rules. This subset preserves FaDe’s DCR and variance improvements but shows higher average FPR deviation, so we treat it as a diagnostic rather than a replacement for the full deterministic protocol. The held-out JSON diagnostic evaluates a new wrapper format against the four predefined formats, asking whether a student trained on text, table, KG, and infobox renderings transfers to an unseen surface wrapper. FaDe gives only modest gains in this unseen-format

Test	Method	N/Settings	Metrics
LLM adjudicator	Base Model	300	Agree. 0.767; Bias 0.227; DCR 0.182; Var. 0.0557
LLM adjudicator	FaDe (Ours)	300	Agree. 0.757; Bias 0.258; DCR 0.244; Var. 0.0457
Held-out JSON	Base Model	8	Bias 0.214; DCR 0.292; Var. 0.0635
Held-out JSON	FaDe (Ours)	8	Bias 0.155; DCR 0.301; Var. 0.0603

Table 18: Adjudicator and out-of-support stress tests. Bias denotes $|\Delta\text{FPR}|$. N/Settings reports response count for LLM-adjudicator rows and format-direction settings for held-out JSON. The held-out JSON rows pair a new JSON rendering with each predefined format in both directions.

Method	N	Rule DCR	LLM DCR	Rule Both kept	LLM Both covered
Base Model	300	0.330	0.200	0.566	0.933
FaDe (Ours)	300	0.383	0.253	0.617	0.934

Table 19: DCR validity audit on 300 sampled Qwen3-8B ConflictBank responses. Rule DCR is the original rule-based Both rate; LLM DCR is the raw Both rate after GPT-4o-mini majority re-adjudication. Rule Both kept is the fraction of rule-Both responses also labeled Both by the LLM adjudicator. LLM Both covered is the fraction of LLM-Both responses recovered by the rule adjudicator.

stress test, indicating that finite-support marginalization should not be assumed to generalize automatically to arbitrary new surface formats.

Table 19 further separates DCR from factual correctness, addressing the main-text warning that higher DCR should not be read as better final-answer accuracy by itself. GPT-4o-mini re-adjudication is more conservative than the rule classifier, lowering absolute DCR for both methods. However, FaDe keeps higher raw LLM-DCR and a larger fraction of rule-based Both labels that remain Both under LLM adjudication. Thus DCR is useful for scalable dual-source coverage comparison, but it remains a noisy diagnostic rather than a standalone correctness metric.

To check whether rule-Both responses are mostly generic hedges, Table 20 audits 100 heterogeneous rule-Both responses per method with LLaMA-3.1-8B-Instruct. The judge checks whether each response substantively covers Claim A and Claim B and flags vague uncertainty without claim content as generic hedging. Most sampled rule-Both responses contain both claim contents, and generic hedging is rare in this subset. This supports interpreting DCR as a claim-coverage diagnostic, while

Method	N	Claim A	Claim B	Dual cov.	Hedge
Base Model	100	97.0	97.0	94.0	0.0
FaDe (Ours)	100	98.0	97.0	95.0	0.0

Table 20: Model-assisted calibration of rule-Both responses on heterogeneous Qwen3-8B ConflictBank outputs. Claim A/B indicate substantive coverage of each claim. Dual cov. requires both claims to be covered and excludes generic hedging. Values are percentages.

Method	Rule label	LLM label			
		A	B	Both	Neither
Base	Pref-A	108	0	1	0
	Pref-B	6	66	3	3
	Both	40	3	56	0
	Neither	6	8	0	0
FaDe	Pref-A	98	0	0	0
	Pref-B	11	57	4	2
	Both	44	0	71	0
	Neither	5	6	1	1

Table 21: Rule-to-LLM adjudicator transition counts on the 300-response Qwen3-8B ConflictBank diagnostic subset. A/B denote single-claim preference labels.

still not treating it as factual correctness or human-verified reasoning.

Table 21 reports the corresponding label-transition counts so that the aggregate DCR change can be traced to concrete label movements. The main failure mode is not answer-string containment; it is that the rule classifier labels many responses as Both whenever both answer strings appear, while GPT-4o-mini often treats the same response as semantically leaning toward one side. This explains why automatic DCR numbers are useful for scalable comparison but too noisy to interpret as factual correctness.

These stress tests define scope rather than add another positive result. The LLM-adjudication sample checks whether the deterministic adjudicator changes the qualitative conclusion, and the JSON-wrapper experiment asks whether a method trained on a finite support transfers to a new wrapper style. The latter result is intentionally reported as a limitation: FaDe should be extended by expanding or learning the format support before being claimed as an arbitrary-format solution. In practice, the intervention support should be treated as a maintained design choice. When a new format is important, users should add renderings involving that format to $\mathcal{A}_{\text{eval}}$, rebuild the corresponding teacher cache, and update the student rather than assuming zero-shot transfer. Online target construction is technically feasible, but it requires multiple teacher passes per example and is more appropriate for

Method	N (side-taking)	Truth-side rate $\uparrow$	DCR on subset $\uparrow$
Base Model	1430	54.7	0.316
FaDe (Ours)	1269	55.9	0.393

Table 22: Side-with-truth accuracy on Qwen3-8B ConflictBank misinformation conflicts. N counts side-taking responses; the truth-side rate is the percentage of these responses that select the verifiably true side.

Method	MuSiQue gold retained $\uparrow$	HotpotQA gold retained $\uparrow$
Base Model	25.6	45.0
FaDe (Ours)	31.9	49.7

Table 23: Gold-answer retention under injected conflict (Qwen3-8B, strict criterion). Values are the percentage of responses that still contain the original gold answer or an alias while omitting the injected counter answer.

high-value samples or periodic offline refreshes than for ordinary single-pass serving.

Factual judgment beyond DCR. DCR measures whether a response covers both conflicting sources, not whether the model resolves the conflict correctly. Tables 22 and 23 therefore measure factual judgment directly on Qwen3-8B. On ConflictBank misinformation conflicts, where one side is verifiably true by construction, FaDe slightly improves the rate at which side-taking responses pick the true side, and DCR on the same subset is also higher. On the MuSiQue and HotpotQA conflicts, where the original gold answer is known, FaDe more often retains the gold answer (or an alias) after counter-evidence injection, under a strict criterion that also requires the injected counter answer to be absent. Reducing format sensitivity thus does not sacrifice factual reliability: marginalization removes only the format-driven component of side-taking, while the gold cross-entropy term anchors the content-driven component.

Teacher-error robustness. A format-marginalized target could in principle smooth incorrect teacher predictions rather than improve factual reliability. Smoothing can only occur where the teacher’s predictions differ across formats: a teacher that is consistently wrong across formats produces little cross-format disagreement, so the mixing weight α_i stays near its minimum and the target falls back to the ordinary teacher distribution, while the gold cross-entropy term and the all-text anchoring path carry factual reliability (Section 3.4). Table 24 tests this directly by stratifying ConflictBank misinformation conflicts

Teacher all-text prediction	N	Truth-side rate Base/FaDe	DCR Base/FaDe
Sides with truth	528	73.3 / 76.4	0.197 / 0.248
Sides against truth	396	25.7 / 22.8	0.131 / 0.194

Table 24: Behavior change stratified by teacher correctness on the Qwen3-8B ConflictBank misinformation subset. Strata are defined by whether the frozen teacher’s all-text prediction sides with the verifiably true claim. Truth-side rate is the percentage of side-taking responses that select the true side.

by whether the frozen teacher’s all-text prediction sides with the truth. When the teacher already sides with the truth, FaDe improves both truth-side selection and dual coverage; when the teacher sides against the truth, truth-side selection stays essentially unchanged within sampling noise while dual coverage rises, i.e., the student more often presents both conflicting sources instead of committing to the teacher’s wrong side. The marginalized target therefore does not simply propagate teacher errors. This is consistent with treating an LLM teacher as a probabilistic, persistently noisy oracle rather than an exact one, a view that also underlies recent work on learning from LLM teachers under fixed label noise (Chen et al., 2024, 2026).

Transfer beyond the predefined support. The held-out JSON rows in Table 18 probe a single unseen wrapper. Tables 25–27 extend this boundary check along three axes on Qwen3-8B. First, Table 25 removes one structured format family entirely from training, dropping all assignments involving it and rebuilding the cache, and then tests on the removed format; all other settings match the main configuration. The held-out format recovers about 92% of the full-support bias reduction on average (almost fully in two of the three arms), with DCR also improving, so the structure-level invariance learned from the remaining formats extends almost undiminished to an unseen member of the trained family. Second, Table 26 repeats the JSON protocol zero-shot on two further wrapper families, YAML and XML, with no support extension, cache rebuild, or retraining. Both show the same trend as JSON: bias falls by 12.3% and DCR improves rather than degrades, so partial transfer to genuinely novel wrappers is not specific to one wrapper. Third, Table 27 applies the unchanged main-table checkpoint zero-shot, under the same controlled format protocol as the main evaluation, to three public real-world conflict datasets: Wiki-

Method	w/o infobox	w/o table	w/o KG
Base Model	0.245 / 0.296	0.299 / 0.319	0.313 / 0.300
FaDe (full support)	0.193 / 0.414	0.246 / 0.430	0.220 / 0.441
FaDe (LOFO)	0.204 / 0.457	0.246 / 0.412	0.223 / 0.441
Gain recovered	79%	100%	97%

Table 25: Leave-one-format-out (LOFO) transfer on Qwen3-8B ConflictBank. Each column removes one structured format from the training support, rebuilds the cache without it, and evaluates on that format; cells report Bias / DCR. Gain recovered is the share of the full-support bias reduction recovered by the LOFO student on the held-out format.

Wrapper	Method	Bias↓	DCR↑	FPR Var.↓
YAML	Base Model	0.300	0.249	0.1075
	FaDe (Ours)	0.263 (−12.3%)	0.271	0.0971
XML	Base Model	0.235	0.285	0.0798
	FaDe (Ours)	0.206 (−12.3%)	0.309	0.0763

Table 26: Zero-shot transfer to two unseen wrapper families on Qwen3-8B ConflictBank, following the held-out JSON protocol of Table 18 with the unchanged checkpoint. Bias denotes $|\Delta\text{FPR}|$; parentheses give the relative bias reduction.

Contradict (Hou et al., 2024), with human-verified contradictions from Wikipedia; QACC (Liu et al., 2025), with conflicting contexts from real search results; and ConflictingQA (Wan et al., 2024), with controversial questions and real web evidence. None of these conflicts appears in training. Format bias falls and dual-source coverage rises on all three sources, with the largest gain on WikiContradict, the closest proxy for naturally occurring conflicts. Overall, transfer is near-complete within the trained structured family, partial for novel wrapper families, and present zero-shot on real-world conflicts. For formats beyond this transfer, the maintenance recipe above is incremental: teacher passes are computed per format assignment, so only the assignments involving a new format require new passes, while existing cached outputs are retained. Retrieval pipelines in other settings, such as multi-step search agents for knowledge-based VQA (Wu et al., 2026), likewise return evidence in heterogeneous surface forms; extending the intervention support to such pipelines is a natural next step.

C Additional Controls, Capability Retention, and Cost

Control map. This section groups mechanism controls, inference-time diagnostics, clean-QA retention checks, and offline-cache cost measurements.

Method	WikiContradict	QACC	ConflictingQA
Base Model	0.164 / 0.143	0.117 / 0.063	0.130 / 0.109
FaDe (Ours)	0.116 / 0.155	0.089 / 0.092	0.108 / 0.134
Bias reduction	−29.3%	−23.9%	−16.9%

Table 27: Zero-shot transfer to three real-world conflict datasets with the unchanged Qwen3-8B FaDe checkpoint; cells report Bias / DCR under the controlled format protocol of the main evaluation.

Method	Bias↓	DCR↑	FPR Var.↓
Base Model	0.276	0.309	0.0847
Vanilla KD (Offline)	<u>0.214</u>	0.346	<u>0.0531</u>
On-Policy KD	0.256	<u>0.408</u>	0.0744
+ Symm. KL	0.257	0.404	0.0758
+ Attn. Bal.	0.250	0.403	0.0726
+ Bidir. Cache	0.261	0.278	0.0774
Attn. Rebal. (Liu et al., 2026)	0.286	0.218	0.0643
Multi-Format KD	0.264	0.319	0.0786
FaDe (Ours)	0.210	0.420	0.0509

Table 28: Mechanism-level controls on Qwen3-8B ConflictBank. Indented rows are focused controls evaluated only in this setting. Lower Bias/FPR Var. is better, and higher DCR indicates greater dual-source coverage. The best result is **bold**, and the second-best result is underlined.

This appendix answers three supporting questions: whether FaDe’s gain comes from format-marginalized targets rather than simpler controls (Appendix C.1), whether all-text anchoring preserves ordinary QA behavior (Appendix C.2), and where the offline compute cost is paid (Appendix C.3). These results are kept separate from the main tables because they use narrower diagnostic settings or practical measurement protocols; they qualify rather than extend the main conclusion, which should be read together with FaDe’s single-pass deployment regime, backbone-dependent retention behavior, and offline cache cost.

C.1 Baseline and Inference-Time Controls

Table 28 tests whether simpler symmetry, attention, or cache variants explain FaDe’s gain in the same Qwen3-8B ConflictBank setting as Table 1. The baseline rows repeat the main setting; the indented rows are focused controls rather than full cross-dataset baselines. They can improve individual metrics but do not match FaDe’s joint bias, variance, and DCR gains.

Table 29 separates same-regime single-pass students from a multi-pass inference-time diagnostic. Format ensembling reduces bias more aggres-

Method	Work	Metrics
Base Model	1 × gen.	Bias 0.276; DCR 0.309; Var. 0.0847
Attn. Rebal. (Liu et al., 2026)	1 × gen. + hook	Bias 0.286; DCR 0.218; Var. 0.0643
On-Policy KD	1 × gen.	Bias 0.256; DCR 0.408; Var. 0.0744
On-Policy KD + Format Ens.	2 × gen.	Bias 0.063; DCR 0.778; Var. 0.0024
FaDe (Ours)	1 × gen.	Bias 0.210 ; DCR 0.420 ; Var. 0.0509

Table 29: Qwen3-8B ConflictBank inference-time upper-bound diagnostic. “Test-time work” counts generation passes per example.

Model	Method	MuSiQue EM	HotpotQA EM
Qwen3-8B	Base Model	27.5	52.5
	On-Policy KD	26.0	51.5
	Multi-Format KD	26.0	52.0
	FaDe (Ours)	27.5	50.0
LLaMA-3.1-8B	Base Model	23.5	54.5
	On-Policy KD	24.0	53.5
	Multi-Format KD	23.0	53.5
	FaDe (Ours)	20.0	47.5

Table 30: Clean QA exact-match retention. EM scores are percentages on all-text clean QA prompts.

sively when repeated generations are allowed, but it multiplies per-query work and yields no reusable debiased student, whereas FaDe pays cache construction and LoRA fine-tuning once and keeps single-pass deployment. The ensembling row is therefore a diagnostic upper bound rather than a same-regime baseline.

C.2 Clean QA Capability Retention

Tables 30 and 31 test non-conflict, all-text QA (EM and F1) on 200 MuSiQue and 200 HotpotQA examples with the main training baselines. Because these prompts contain no constructed conflict or heterogeneous evidence pair, they probe whether the anchoring path of Section 3.3 preserves ordinary QA behavior after debiasing. Qwen3-8B preserves clean QA F1 under FaDe, while LLaMA-3.1 shows a 4.0-point drop; we treat this as a targeted retention check rather than a substitute for broad capability benchmarks.

Table 32 repeats the Base-vs-FaDe comparison on 400 examples per dataset. The larger sample strengthens the Qwen3-8B result and reduces, but does not remove, the LLaMA-3.1 degradation, so retention remains empirical and backbone-dependent.

Anchoring strength on LLaMA-3.1. Since Qwen3-8B improves under the same objective, the

Model	Method	MuSiQue F1	HotpotQA F1	Avg./ Δ
Qwen3-8B	Base Model	40.5	68.2	54.4 / 0.0
	On-Policy KD	38.9	67.9	53.4 / -1.0
	Multi-Format KD	39.1	68.7	53.9 / -0.5
	FaDe (Ours)	42.7	66.7	54.7 / +0.3
LLaMA-3.1-8B	Base Model	34.1	69.2	51.7 / 0.0
	On-Policy KD	36.1	68.4	52.3 / +0.6
	Multi-Format KD	35.1	68.1	51.6 / -0.1
	FaDe (Ours)	31.7	63.7	47.7 / -4.0

Table 31: Clean QA F1 retention. Δ vs. each backbone’s base model.

Model	Method	MuSiQue EM	MuSiQue F1	HotpotQA EM	HotpotQA F1	Avg. F1/ Δ
Qwen3-8B	Base Model	26.5	39.5	53.8	69.0	54.2 / 0.0
	FaDe (Ours)	28.2	41.6	56.8	71.0	56.3 / +2.1
LLaMA-3.1-8B	Base Model	23.0	34.2	54.5	68.4	51.3 / 0.0
	FaDe (Ours)	21.5	32.7	52.2	65.9	49.3 / -2.0

Table 32: Expanded clean QA retention check. Scores are percentages on 400 all-text examples from each dataset. Δ is relative to the base model within each backbone.

Configuration	Clean QA EM / F1 (Δ F1)	Bias $\downarrow$	DCR $\uparrow$
Base Model	38.8 / 51.3 (0.0)	0.103	0.331
FaDe (default)	36.9 / 49.3 (-2.0)	0.057	0.503
FaDe (strengthened)	38.1 / 52.1 (+0.8)	0.058	0.487

Table 33: Anchoring-strength mitigation on LLaMA-3.1-8B. Default and strengthened denote the anchoring configuration; all other settings are identical. Clean QA reports average EM / F1 over 400 all-text MuSiQue and HotpotQA examples (protocol of Table 32); Δ F1 is relative to the base model. Bias and DCR are measured on ConflictBank.

LLaMA-3.1 drop is backbone-dependent rather than an inherent cost of format marginalization. The anchoring strength is configurable through λ_{anc} in Eq. (9) and the share of all-text anchoring samples seen during training (defaults in Table 4), so a natural hypothesis is that the default anchoring is too weak for LLaMA-3.1. Table 33 retrains FaDe on LLaMA-3.1-8B with a strengthened anchoring configuration, keeping everything else identical to the main setting, and reports clean-QA retention (protocol of Table 32) together with ConflictBank debiasing. The drop disappears, with clean-QA F1 slightly above the base model, while the debiasing gains stay essentially intact. Retention therefore depends on the anchoring configuration, which we recommend tuning per backbone; broader capability benchmarks remain a future check.

Method	Offline work	Runtime
Vanilla KD (Offline)	1 pass; 165M cache	Cache 21m; train 5h52m; eval 36m
On-Policy KD	1 pass; 164M cache	Train 1h28m; eval 2h17m
Multi-Format KD	12 passes; 2.3G cache	Cache 1h42m; train 6h18m; eval 53m [†]
FaDe (Ours)	7 passes; 329M cache	Cache 1h41m ; train 5h58m ; eval 1h39m

[†]Multi-Format evaluation was split across two GPUs; the value is elapsed wall time.

Table 34: Computational overhead on the main ConflictBank setting, measured from completed Qwen3-8B logs.

C.3 Computational Cost and Overhead

Hardware and compute budget. All experiments were run on a single Linux server with two NVIDIA A100-SXM4-80GB GPUs, an Intel Xeon Gold 6330 CPU, and 125 GiB RAM. Unless otherwise noted, each job used one GPU; the Multi-Format KD evaluation in Table 34 was split across both GPUs. The final reported local experiments used approximately 350 A100 GPU-hours, estimated from completed wall-clock logs; this excludes exploratory/debug runs, CPU-only post-processing, and external GPT-4o-mini API adjudication.

FaDe moves teacher computation to offline cache construction. Table 34 reports teacher passes and cache size (offline work) together with representative Qwen3-8B wall-clock times for cache, training, and evaluation jobs (runtime); timings indicate order of magnitude rather than a controlled systems benchmark, and the table shows where FaDe pays for debiasing rather than claiming a hardware-optimized implementation.

Unlike format ensembling, FaDe keeps single-pass inference, and storing only the top- K mixed target keeps its cache smaller than full Multi-Format KD. Two orthogonal levers could reduce the offline cost further: input-side compression, as used for visual tokens in multimodal code and document understanding (Tang et al., 2026a,b), would shorten each teacher pass, since passes dominate cache construction and scale with rendered evidence length; and inference-time routing between models of different sizes (Ye et al., 2026b) is an alternative way to control cost, which FaDe sidesteps by fixing a single student and paying the cost once offline.