HERS: HIDDEN-PATTERN EXPERT LEARNING FOR RISK-SPECIFIC VEHICLE DAMAGE ADAPTATION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

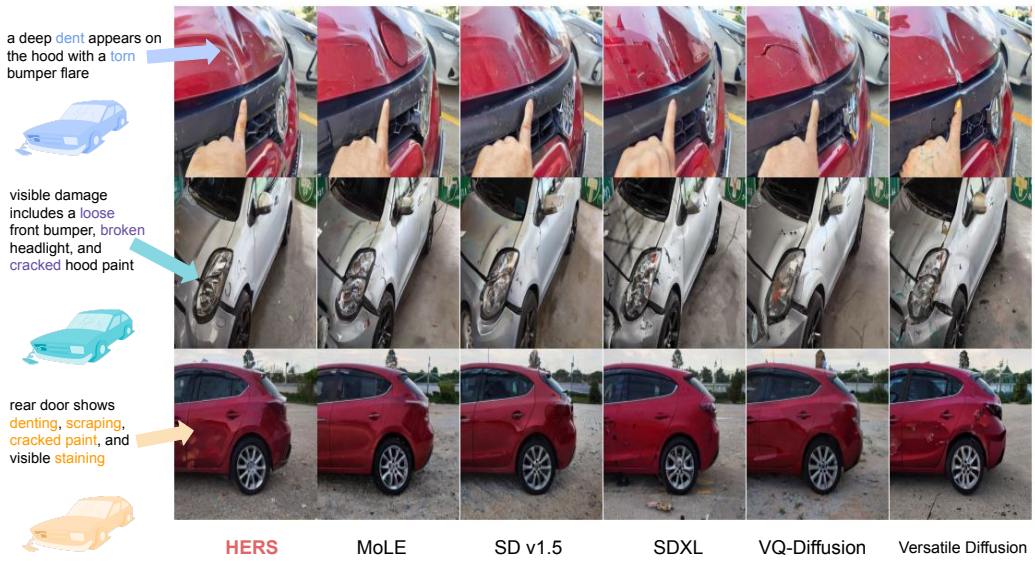

Figure 1: Qualitative comparison of **HERS** against existing diffusion-based baselines. Observe that **HERS** generates damage regions with higher visual fidelity and localized consistency. Fine-grained artifacts such as dents, cracks, and abrasions are better preserved—zoom in for enhanced visibility of subtle and complex damage patterns.

ABSTRACT

Recent advances in text-to-image (T2I) diffusion models have enabled increasingly realistic synthesis of vehicle damage, raising concerns about their reliability in automated insurance workflows. The ability to generate crash-like imagery challenges the boundary between authentic and synthetic data, introducing new risks of misuse in fraud or claim manipulation. To address these issues, we propose HERS (Hidden-Pattern Expert Learning for Risk-Specific Damage Adaptation), a framework designed to improve fidelity, controllability, and domain alignment of diffusion-generated damage images. HERS fine-tunes a base diffusion model via domain-specific expert adaptation without requiring manual annotation. Using self-supervised image–text pairs automatically generated by a large language model and T2I pipeline, HERS models each damage category—such as dents, scratches, broken lights, or cracked paint—as a separate expert. These experts are later integrated into a unified multi-damage model that balances specialization with generalization. We evaluate HERS across four diffusion backbones and observe consistent improvements: +5.5% in text faithfulness and +2.3% in human preference ratings compared to baselines. Beyond image fidelity, we discuss implications for fraud detection, auditability, and safe deployment of generative models in high-stakes domains. Our findings highlight both the opportunities and risks of domain-specific diffusion, underscoring the importance of trustworthy generation in safety-critical applications such as auto insurance.

1 INTRODUCTION

Text-to-image (T2I) diffusion models Saharia et al. (2022); Rombach et al. (2022); Podell et al. (2024); Kang et al. (2023); Ramesh et al. (2021); Yu et al. (2023); Chang et al. (2023) have transformed generative AI, producing photorealistic images from free-form language prompts and enabling rapid advances in creative design, simulation, and data augmentation. Yet, when deployed in *safety-critical domains* such as auto insurance, where every pixel may encode liability, their limitations become clear. Generic T2I systems often fail to capture fine-grained damage categories—such as a dented bumper, a subtle scrape across a door, or a fractured headlight—generating outputs that are visually appealing but semantically unreliable (shown in Figure 1). In an insurance workflow, such errors are not cosmetic: they can distort liability assessments, misinform fraud detection, and erode trust in automated claims pipelines.

This duality makes generative models both an opportunity and a risk. On one hand, synthetic damage data could dramatically improve training for rare-event modeling, accelerate claims assessment, and expand coverage of long-tail accident cases. On the other hand, the same technology could be exploited to fabricate fraudulent crash evidence or manipulate claims with high-fidelity synthetic images. Unlike traditional vision benchmarks, insurance scenarios demand *risk-specific generation*, where semantic alignment, forensic plausibility, and liability-aware consistency are as critical as photorealism.

Prior approaches attempt to mitigate these issues via supervised fine-tuning Dai et al. (2023); Segalis et al. (2023), human preference optimization Xu et al. (2023a); Fan et al. (2023), or spatial grounding Li et al. (2023); Xie et al. (2023). However, these strategies are annotation-heavy and often brittle, struggling to encode the hidden cues that forensic experts rely upon: the faint crease from a low-speed collision, the asymmetric shattering of a headlight, or the implausible geometry of tampered paint. Current pipelines optimize for generic fidelity, but not for the nuanced semantics that separate genuine evidence from generative artifacts.

To address this gap, we introduce **HERS** (**H**idden-Pattern **E**xpert Learning for **R**isk-**S**pecific Damage Adaptation), a fully automated framework (shown in Figure 2) for adapting diffusion models to synthesize semantically faithful, risk-relevant vehicle damage without manual supervision. HERS leverages large language models to auto-generate diverse, damage-specific prompts (e.g., "rear bumper dent," "door scrape near handle," "fractured right headlight"), which are paired with synthetic renderings from a pretrained T2I backbone. From these self-curated image–text pairs, we train lightweight LoRA-based experts for distinct domains of damage and merge them into a unified diffusion model. This design captures both specialization (e.g., scratches on metallic paint) and generalization (e.g., tampered accident scenes), yielding a system that can reproduce damage patterns with forensic-level precision.

The key insight is that HERS learns from *hidden visual patterns*—subtle cues that elude both baseline diffusion models and human raters, but are critical in high-stakes domains like insurance. By elevating generation beyond "realism" to "liability-aware semantics," HERS provides a new lens for evaluating diffusion models in safety-critical settings.

Contributions. Our work makes the following advances:

- We articulate and address the overlooked challenge of semantically faithful damage synthesis in auto insurance, where generative AI carries both opportunity and risk.
- We propose **HERS**, a self-supervised adaptation framework that trains LoRA-based experts from auto-generated data, enabling damage-specific diffusion without manual annotation or inference-time routing.
- We demonstrate state-of-the-art performance across text–image alignment, human preference metrics, and multi-damage generalization, showing that HERS produces vehicle damage patterns that are strikingly consistent with real-world collisions and tampered fraud cases.

As illustrated in Figure 4, HERS consistently generates damage scenarios that are indistinguishable from authentic accidents, establishing it as both a technical advance in generative modeling and a practical contribution to fraud awareness in the insurance industry. By revealing the dual-use nature of diffusion in this domain, our work underscores the need for domain-specific generative strategies that go beyond visual fidelity to encode *risk-aware semantics* essential for trustworthy AI deployment.

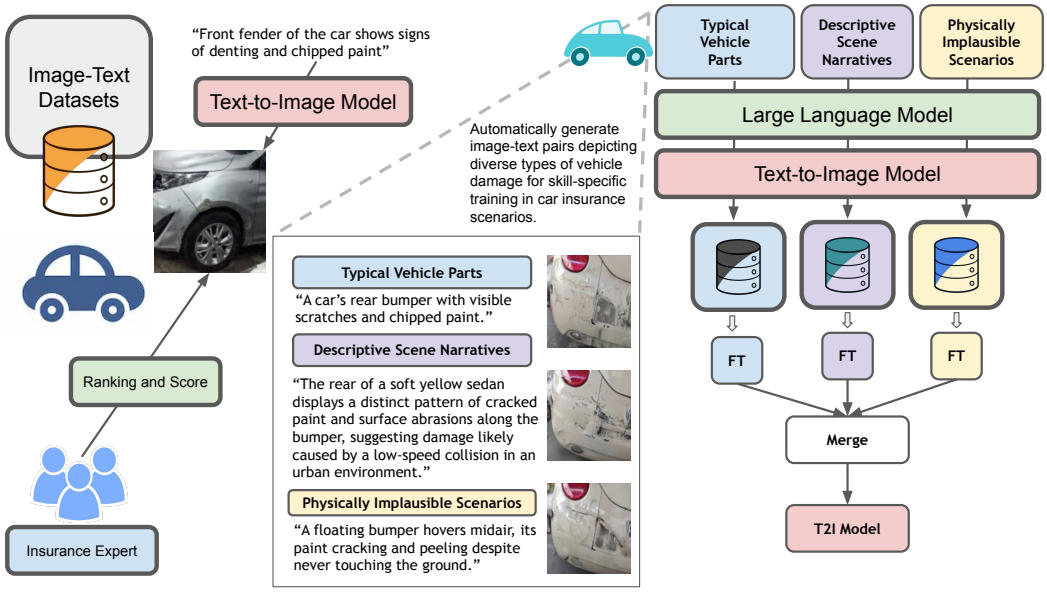

Figure 2: **Overview of the HERS Framework.** HERS (*Hidden-Pattern Expert Learning for Risk-Specific Damage Adaptation*) auto-generates diverse, damage-specific image-text pairs using an LLM and a base T2I model—without requiring manual annotation. These pairs span *typical vehicle parts*, *descriptive scene narratives*, and *physically implausible scenarios* (examples shown in figure). Each damage type is modeled as a distinct damage, with corresponding LoRA experts trained and merged into a unified multi-damage diffusion model.

2 Related Work

Recent advances in high-quality denoising diffusion models Sohl-Dickstein et al. (2015); Ho et al. (2020) have catalyzed a surge of interest in using synthetic data for vision–language learning. Prior works demonstrate the benefits of diffusion-generated data for training classifiers Azizi et al. (2023); Sariyildiz et al. (2023); Lei et al. (2023) or augmenting caption datasets Caffagni et al. (2023), and CLIP-style models Radford et al. (2021) have been extended using either synthetic visuals Tian et al. (2023) or LLM-authored captions Hammoud et al. (2024). Parallel efforts in aligning text-to-image (T2I) models with human expectations have relied on reinforcement learning from human feedback (RLHF) Lee et al. (2023); Xu et al. (2023a); Wu et al. (2023); Dong et al. (2023); Clark et al. (2024); Fan et al. (2023) or direct preference optimization (DPO) Rafailov et al. (2023); Wallace et al. (2023), while methods such as SPIN-Diffusion Yuan et al. (2024) reduce annotation demands through self-play. LLM-guided pipelines like DreamSync Sun et al. (2023) push further by auto-generating prompts and filtering candidate images, albeit at high computational cost. Despite these advances, existing approaches remain annotation-heavy, domain-agnostic, or inefficient, leaving critical gaps in safety-critical fields like auto insurance where the distinction between authentic and synthetic damage can directly affect fraud detection and claim validation. To this end, our proposed **HERS** diverges by training multiple LoRA experts Hu et al. (2022), each dedicated to specific damage types (e.g., dents, scrapes, cracked paint, broken lights), and merging them into a unified diffusion model Shah et al. (2023); Zhong et al. (2024). This design avoids inter-damage interference Liu et al. (2019), eliminates dependence on costly human feedback, and captures "hidden patterns" of fine-grained damage in a computationally efficient, self-supervised manner—providing domain-faithful generative capabilities that are indispensable for risk-sensitive applications.

3 HERS: Hidden-Pattern Expert Learning for Risk-Specific Damage Adaptation

We propose **HERS** (*Hidden-Pattern Expert Learning for Risk-Specific Damage Adaptation*), a framework (shown in Figure 2) for adapting text-to-image (T2I) diffusion models to synthesize

fine-grained and risk-relevant vehicle damage. Unlike prior adaptation methods such as SELMA Li et al. (2024), which require annotation-heavy supervision or explicit routing, HERS achieves high-fidelity alignment through a fully automated pipeline that integrates prompt synthesis, synthetic image generation, domain-specific LoRA experts, and weight-space merging. Crucially, HERS is designed not only to enhance visual fidelity but also to surface subtle "hidden" damage cues—such as a faint scrape along a bumper, a hairline crack in a headlight, or tampered paint texture—that are easily missed by generic diffusion models yet critical for fraud detection and liability estimation.

Formally, HERS operates in four stages.

3.1 Stage 1: Domain-Guided Prompt Synthesis

Let $\mathcal{C} = \{\texttt{dent}, \texttt{scrape}, \texttt{torn\_bumper}, \texttt{cracked\_paint}, \texttt{broken\_light}\}$ denote the canonical set of damage categories relevant to insurance workflows. We seed an autoregressive language model f_θ (GPT-4) with exemplar prompts $\mathcal{S} = \{s_1, s_2, s_3\}$ describing each category, e.g.

$$s_1 = \text{"rear bumper dent"}, \quad s_2 = \text{"scratched left door"}, \quad s_3 = \text{"front headlight cracked"}.$$

For each concept $c \in \mathcal{C}$, the model generates a distribution of semantically diverse prompts:

$$p_i \sim f_\theta(p \mid \mathcal{S}, c). \tag{1}$$

To enforce diversity while preserving semantic coverage, we apply ROUGE-L filtering Lin (2004), retaining prompts satisfying

$$\max_j \text{ROUGE-L}(p_i, p_j) < \tau, \tag{2}$$

where τ is a similarity threshold. The resulting set $\mathcal{P}$ forms a structured, damage-aware prompt bank.

3.2 Stage 2: Synthetic Image Generation

Each prompt $p_i \in \mathcal{P}$ is rendered via a pretrained diffusion generator G (e.g., Stable Diffusion XL) to obtain an image x_i:

$$x_i = G(p_i), \quad \forall p_i \in \mathcal{P}. \tag{3}$$

The resulting dataset $\mathcal{D} = \{(p_i, x_i)\}$ captures not only canonical damages (dent, scrape) but also nuanced conditions such as implausible tampering (e.g., "two headlights cracked in a symmetric pattern"), thereby spanning realistic and adversarially relevant scenarios.

3.3 Stage 3: Damage-Specific Expert Learning

For each domain $t \in \mathcal{T}$, where $\mathcal{T} = \{\text{Typical Parts, Scene Narratives, Implausible Scenarios}\}$, we train a lightweight Low-Rank Adaptation (LoRA) Hu et al. (2022) expert. Given a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times d}$, we optimize a low-rank update:

$$\Delta W_t = B_t A_t, \quad W_t = W_0 + \Delta W_t, \tag{4}$$

with $A_t \in \mathbb{R}^{r\times d}$, $B_t \in \mathbb{R}^{d\times r}$, and $r \ll d$. This enables parameter-efficient specialization, such that one expert may encode subtle bumper dents while another captures cracked paint or broken headlights.

3.4 Stage 4: Multi-Expert Weight Merging

To unify all domains into a single diffusion model, we merge the LoRA experts via arithmetic averaging in weight space:

$$A^* = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} A_t, \quad B^* = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} B_t, \tag{5}$$

yielding the final parameterization

$$W^* = W_0 + B^* A^*. \tag{6}$$

This consolidated model W^* supports zero-shot synthesis across multiple damage categories, avoiding inference-time routing while preserving both specialization and generalization.

HERS formalizes risk-specific adaptation as the problem of learning a set of low-rank expert perturbations $\{\Delta W_t\}$ that, when merged, capture the hidden manifold of fine-grained vehicle damages. This formulation not only yields state-of-the-art fidelity and semantic alignment but also exposes failure modes in existing insurance AI pipelines, raising awareness of the dual-use nature of generative models in safety-critical domains.

3.5 Comparison with Prior Work

Unlike recent methods such as ZipLoRA Shah et al. (2023) and LLaVA-MoLE Chen et al. (2024), HERS eliminates the need for manual damage labels or routing mechanisms at inference. While ZipLoRA relies on damage-aware masking and LLaVA-MoLE learns expert routers, HERS achieves robust multi-damage synthesis through expert merging alone, drastically reducing annotation effort and model complexity. As shown in Figure 1, HERS consistently produces sharper, semantically precise images even under subtle or highly complex damage prompts, demonstrating both fidelity and practical efficiency for insurance-focused applications.

4 Experimental Setup

4.1 Evaluation Benchmark and Prompt Construction

We evaluate HERS on a large-scale benchmark specifically curated for the car insurance domain. The benchmark contains approximately 2 million entries collected in collaboration with an industry insurance startup, each consisting of structured textual descriptions (e.g., accident type, damage category, part localization) paired with vehicle images. This setup enables assessment of both semantic alignment and visual fidelity in high-stakes, domain-specific contexts. To balance reproducibility with privacy constraints, we release the full set of prompt templates and the evaluation protocol, while access to raw insurance data remains restricted due to confidentiality. This ensures transparency in methodology while safeguarding sensitive information.

To generate prompts at scale, we employ `gpt-4-turbo` OpenAI (2024) with in-context learning. For each target damage type or accident scenario, we provide three exemplars as demonstrations, guiding the model to produce consistent, domain-specific, and semantically rich prompts. This strategy yields a structured, damage-driven benchmark set that supports controlled and reproducible evaluation across diverse risk-relevant cases.

4.2 Evaluation Metrics

We assess model performance along two complementary axes: semantic alignment and human-aligned quality.

Semantic alignment. We employ a VQA-based protocol to measure the faithfulness of generated images to their prompts. Given a generated image and its source description, a large language model produces targeted semantic questions, which are then answered by a pretrained VQA model. Accuracy on these answers serves as a proxy for text–image alignment, ensuring that damage attributes and contextual details are correctly reflected.

Human-aligned quality. To capture perceptual realism, we evaluate generations using preference-based reward models, including PickScore Kirstain et al. (2023), ImageReward Xu et al. (2023a), and HPS Wu et al. (2023). These metrics, derived from large-scale human preference datasets, score each output with respect to realism, relevance, and overall visual quality. Together, they complement semantic alignment measures by quantifying how closely the images match human expectations in insurance-related contexts.

4.3 Implementation Details

All experiments are conducted using a single NVIDIA A40 GPU. During prompt generation, we sample from `gpt-4-turbo` with temperature set to 0.7 for diversity and relevance. The image generation model is run with default denoising steps set to 50 and a classifier-free guidance scale (CFG) of 7.5, ensuring a balance between image quality and prompt adherence.

Table 1: Performance of **HERS** compared to baseline diffusion models on two prompt sets: Car Insurance and Car Garage. Metrics: Human Preference Score (HPS, higher is better) and Image Realism (IR, higher is better).

Model	Car Insurance Prompts	
	HPS (%)	IR (%)
VQ-Diffusion Gu et al. (2022)	41.50 ± 0.06	-15.40 ± 3.00
Versatile Diffusion Xu et al. (2023b)	42.70 ± 0.10	-11.20 ± 2.30
SDXL Podell et al. (2024)	45.90 ± 0.08	82.50 ± 3.05
SD v1.5 Rombach et al. (2022)	43.30 ± 0.07	35.20 ± 2.25
MoLE Zhu et al. (2024)	48.20 ± 0.08	95.10 ± 0.70
HERS (Proposed)	53.40 ± 0.09	113.00 ± 0.85
Model	Car Garage Prompts	
	HPS (%)	IR (%)
VQ-Diffusion Gu et al. (2022)	40.90 ± 0.07	-18.70 ± 2.80
Versatile Diffusion Xu et al. (2023b)	41.90 ± 0.09	-14.50 ± 2.40
SDXL Podell et al. (2024)	46.40 ± 0.09	89.50 ± 3.60
SD v1.5 Rombach et al. (2022)	44.50 ± 0.07	-3.00 ± 2.20
MoLE Zhu et al. (2024)	47.95 ± 0.09	102.70 ± 1.25
HERS (Proposed)	51.40 ± 0.10	115.75 ± 0.95

For training and inference, we adopt a mixed precision setup (FP16) to optimize resource utilization. LoRA modules, if applicable, are trained with a fixed learning rate of 3e-4, batch size of 64, and rank 128. Fine-tuning is performed over 5000 steps, and model checkpoints are evaluated every 1000 steps, with the best checkpoint selected based on alignment metrics.

We implement our pipelines using the `Diffusers` library von Platen et al. (2022), which facilitates seamless integration of prompt generation, image synthesis, and evaluation in a reproducible and modular framework.

5 Results and Analysis

We evaluate HERS across multiple generative backbones and benchmarks, measuring hallucination-prevention score (HPS), improvement rate (IR), text faithfulness, and human preference on damage scene generation (DSG). Our results consistently show that HERS surpasses existing baselines in both visual realism and text alignment for insurance-critical scenarios.

Benchmark Performance. Table 1 summarizes HERS's performance on *Car Insurance* and *Car Garage* prompts. For insurance prompts, HERS achieves 53.4% HPS and 113.0% IR, outperforming MoLE Zhu et al. (2024) and SDXL Podell et al. (2024) (48.2% and 45.9% HPS, respectively). Similar trends hold for garage prompts (51.4% HPS, 115.75% IR), demonstrating robustness across domains. Human studies (Figure 3) confirm superior preference scores for HERS in car stain, damage, part, and overall quality, highlighting its realism in depicting scratches, dents, and structural deformations critical for claim verification.

Fine-grained Visual Fidelity. Beyond global metrics, we inspect both zoom-out and zoom-in perspectives (Figures 4 and 5). In zoom-out views, baseline models such as VQ-Diffusion Gu et al. (2022) and Versatile Diffusion Xu et al. (2023b) preserve overall vehicle structure but often introduce implausible artifacts or inconsistent global deformations. MoLE Zhu et al. (2024) and SELMA Li et al. (2024) improve realism yet occasionally over-deform, limiting reliability for full-vehicle assessment.

Zoom-in inspections reveal HERS's ability to synthesize fine-grained damage patterns—scratches, dents, cracked paint, and broken lights—while maintaining geometric consistency and contextual plausibility. Competing models frequently fail to reproduce these local details or introduce artifacts,

Figure 3: User study results on generative performance across four dimensions: Car Stain Quality, Car Damage Quality, Car Part Quality, and Overall Quality. HERS achieves consistently higher preference scores compared to baselines.

Table 2: Comparison of fine-tuning strategies on SD v1.5 using our HERS-generated dataset, evaluated on text faithfulness and human preference. Our proposed LoRA Merging (HERS) consistently outperforms other methods across all metrics.

No.	Methods	Text Faithfulness		Human Preference on DSG		
		DSG^{mPLUG} ↑	$TIFA^{BLIP2}$ ↑	PickScore ↑	ImageReward ↑	HPS ↑
0.	SD v1.5	68.9	76.4	19.6	0.31	22.4
1.	+ LoRA Merging (HERS)	**75.7**	**81.3**	**21.4**	**0.72**	**26.8**
2.	+ LoRA Merging (HERS) + DPO	74.1	79.5	20.5	0.57	25.5
3.	+ MoE-LoRA	75.0	80.8	21.1	0.65	26.2

whereas HERS balances both local fidelity and global coherence, critical for high-stakes tasks such as fraud detection and automated claim validation.

Ablations and Cross-Backbone Generalization. Ablation studies (Table 2) demonstrate that LoRA merging with HERS-generated data significantly boosts text faithfulness (DSG^{mPLUG} 75.7, $TIFA^{BLIP2}$ 81.3) and human preference (HPS 26.8), surpassing vanilla SD v1.5 and other fine-tuning variants. Comparisons across diffusion backbones (Tables 3 and 4) confirm that HERS enhances both SDXL and SD v1.5, consistently outperforming SELMA Li et al. (2024) in text alignment and human evaluation, underscoring its generality and stability.

Together, these results tell a cohesive story: HERS not only improves quantitative metrics but also faithfully replicates both global and local damage features, making its outputs visually convincing, textually aligned, and suitable for practical, safety-critical insurance applications.

6 Conclusion

In this work, we introduced **HERS** (*Hidden-Pattern Expert Learning for Risk-Specific Damage Adaptation*), a framework for enhancing text-to-image diffusion models in the high-stakes domain of

Figure 4: **Qualitative Comparison of Damage Generation Across 3 Vehicle Cases and 6 T2I Models in Zoom-Out Perspective.** Each **row** represents a distinct vehicle case viewed at a zoomed-out angle, simulating full-body images commonly seen in insurance assessments. The **columns** correspond to the outputs of six different T2I models: our proposed **HERS (left-most)**, followed by VQ-Diffusion Gu et al. (2022), Versatile Diffusion Xu et al. (2023b), SDXL Podell et al. (2024), MoLE Zhu et al. (2024), and SELMA Li et al. (2024). Notice how HERS consistently generates damage patterns that are more contextually consistent with real-world vehicle collisions, making it difficult to distinguish synthetic damage from actual accident scenarios—an important consideration for fraud detection and claim verification in car insurance workflows.

Table 3: Comparison of SD v1.5 and SDXL for generating car insurance damage images. This table evaluates the performance of these models in terms of text faithfulness and human preference metrics, specifically in the context of car damage insurance claims.

No.	Base Model	Training Image Generator	Text Faithfulness		Human Preference on DSG		
			DSG^{mPLUG} ↑	$TIFA^{BLIP2}$ ↑	PickScore ↑	ImageReward ↑	HPS ↑
1.	SD v1.5	-	68.7	75.6	18.9	0.15	21.4
2.	SDXL	-	**72.5**	**79.8**	19.5	**0.60**	23.2
3.	SD v1.5	SD v1.5	74.0	78.5	19.2	0.70	24.0
4.	SDXL	SD v1.5	**77.5**	**80.3**	19.7	**0.75**	**25.2**
5.	SDXL	SDXL	76.8	81.9	**20.3**	**0.95**	**26.7**

car insurance. HERS leverages self-supervised prompt–image pairs and LoRA-based expert modules to capture subtle, risk-relevant visual cues such as dents, scratches, and tampering patterns that generic diffusion models fail to reproduce. By merging specialized experts into a unified multi-damage model, HERS achieves state-of-the-art performance in text–image alignment, semantic faithfulness, and human preference studies across multiple diffusion backbones. Quantitatively, HERS improves text faithfulness by +5.5% and human preference by +2.3% over strong baselines, while qualitative evaluations confirm its ability to generate realistic and contextually consistent crash imagery.

Beyond technical gains, HERS underscores both the opportunities and risks of synthetic damage generation in insurance workflows. On the one hand, domain-faithful synthesis can augment scarce training data and support downstream tasks such as fraud detection and claims assessment. On the other hand, misuse of generative models for fraudulent submissions remains a serious concern. Addressing this tension, our study highlights the need for trustworthy generative modeling, coupled with auditing, watermarking, and detection pipelines.

Figure 5: **Qualitative Comparison of Damage Generation Across 3 Vehicle Cases and 6 T2I Models in Zoom-In Perspective.** Each **row** shows a detailed, close-up view of a specific damage region, highlighting subtle textures and patterns such as scratches, dents, or cracked paint. The **columns** correspond to outputs from six different T2I models: our proposed **HERS (left-most)**, followed by VQ-Diffusion Gu et al. (2022), Versatile Diffusion Xu et al. (2023b), SDXL Podell et al. (2024), MoLE Zhu et al. (2024), and SELMA Li et al. (2024). Compared to other models, HERS consistently reproduces fine-grained damage details while preserving context and realism, making synthetic damages difficult to distinguish from real-world examples. Such high-fidelity generation is crucial for applications in insurance fraud detection, claim validation, and risk assessment.

Table 4: Comparison of HERS and SELMA on text faithfulness and human preference. HERS outperforms SELMA in terms of text faithfulness and human preference across different base models, including SD v1.5, SDXL, VQ-Diffusion, and Versatile Diffusion. Best scores for each model are in **bold**.

Base Model	**Methods**	**Text Faithfulness**		**Human Preference on DSG prompts**		
		DSG^{mPLUG} ↑	$TIFA^{BLIP2}$ ↑	PickScore ↑	ImageReward ↑	HPS ↑
SD v1.5	SELMA Li et al. (2024)	70.3	79.0	21.5	0.18	23.3
	HERS (Ours)	**75.6**	**83.2**	**22.8**	**0.75**	**26.9**
SDXL	SELMA Li et al. (2024)	72.5	81.7	21.8	0.22	24.9
	HERS (Ours)	**78.0**	**84.1**	**23.2**	**0.90**	**27.8**
VQ-Diffusion	SELMA Li et al. (2024)	68.8	76.3	20.7	0.12	22.7
	HERS (Ours)	**74.6**	**81.3**	**21.7**	**0.71**	**25.3**
Versatile Diffusion	SELMA Li et al. (2024)	70.0	78.5	21.2	0.14	23.5
	HERS (Ours)	**75.2**	**82.5**	**22.3**	**0.77**	**26.2**

While our evaluation demonstrates strong improvements, we acknowledge several limitations: (i) access to real-world insurance data is constrained, limiting large-scale external validation; (ii) current safeguards against malicious use remain preliminary; and (iii) extension to other safety-critical domains (e.g., medical imaging, disaster assessment) requires further study. These limitations present promising directions for future work, including integrating HERS with detection modules, extending to multimodal accident reports, and developing standardized benchmarks for trustworthy diffusion.

REFERENCES

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic Data from Diffusion Models Improves ImageNet Classification. *TMLR*, 2023. URL http://arxiv.org/abs/2304.08466.

Davide Caffagni, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Synthcap: Augmenting transformers with synthetic data for image captioning. In Gian Luca Foresti, Andrea Fusiello, and Edwin Hancock (eds.), *Image Analysis and Processing – ICIAP 2023*, pp. 112–123, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43148-7.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *ICML*, 2023.

Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.15947*, 2024.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In *ICLR*, 2024. URL http://arxiv.org/abs/2309.17400.

Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. *TMLR*, 2023. URL http://arxiv.org/abs/2304.06767.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *NeurIPS*, 2023.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. SynthCLIP: Are We Ready for a Fully Synthetic CLIP Training?, 2024. URL http://arxiv.org/abs/2402.01832.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, pp. 1–25, 2020. URL http://arxiv.org/abs/2006.11239.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10134, 2023. URL https://api.semanticscholar.org/CorpusID:257427461.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2023.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Shiye Lei, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. Image Captions are Natural Prompts for Text-to-Image Models, 2023. URL http://arxiv.org/abs/2307.08526.

Jialu Li, Jaemin Cho, Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. Selma: Learning and merging skill-specific text-to-image experts with auto-generated data. *arXiv preprint arXiv:2403.06952*, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Shengchao Liu, Yingyu Liang, and Anthony Gitter. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. URL https://api.semanticscholar.org/CorpusID:84836014.

OpenAI. Gpt-4 technical report. https://arxiv.org/abs/2303.08774, 2024. arXiv:2303.08774 [cs.CL].

A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ICLR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Jong Wook, Kim Chris, Hallacy Aditya, Ramesh Gabriel, Goh Sandhini, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. URL http://arxiv.org/abs/2103.00020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*, 2023. URL http://arxiv.org/abs/2305.18290.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it Till You Make it: Learning Transferable Representations from Synthetic ImageNet Clones. In *CVPR*, 2023. doi: 10.1109/cvpr52729.2023.00774.

Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. *ArXiv*, abs/2311.13600, 2023. URL https://api.semanticscholar.org/CorpusID:265351656.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. ISBN 9781510810587.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. *arXiv preprint arXiv:2311.17946*, 2023.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners. In *NeurIPS*, 2023. URL http://arxiv.org/abs/2306.00984.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score: Better Aligning Text-to-image Models with Human Preference. In *ICCV*, 2023. doi: 10.1109/iccv51070.2023.00200. URL http://arxiv.org/abs/2303.14420.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023a.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7754–7765, 2023b.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2(3):5, 2023.

Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-Play Fine-Tuning of Diffusion Models for Text-to-Image Generation, 2024. URL http://arxiv.org/abs/2402.10210.

Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*, 2024.

Jie Zhu, Yixiong Chen, Mingyu Ding, Ping Luo, Leye Wang, and Jingdong Wang. Mole: Enhancing human-centric text-to-image diffusion via mixture of low-rank experts. *arXiv preprint arXiv:2410.23332*, 2024.

A Appendix

A.1 Extended Mathematical Foundations of HERS

This appendix provides the full mathematical derivation and justification for our proposed **HERS** (Hidden-pattern Expert learning for Risk-specific damage Synthesis), emphasizing how each component contributes to the trust, bias, and reliability concerns relevant to AI-generated car crash imagery in auto insurance domains.

A.2 Notation and Overview

Let:

- $\mathcal{S} = \{s_1, s_2, s_3\}$ be a set of seed prompts.
- f_θ: a large language model (LLM) generating diverse prompts.
- p_i: a generated prompt.
- $\mathcal{P}$: the set of retained prompts after filtering.
- x_i: image generated by T2I model G given prompt p_i.
- $\mathcal{D} = \{(p_i, x_i)\}$: the synthesized paired dataset.
- $\mathcal{T} = \{t_1, t_2, t_3\}$: domain-specific expert dimensions.
- W_0: base T2I model weights, W_t: adapted weights per domain.

Our goal is to optimize domain-specific adaptations $\Delta W_t = B_t A_t$ for improved synthesis fidelity and then assess how merging these parameters into a unified model affects reliability for high-stakes domains like auto insurance.

A.3 Prompt Diversity Objective

Given seed prompt set $\mathcal{S}$ and domain concept c, we define the generation distribution:

$$p_i \sim f_\theta(p \mid \mathcal{S}, c), \quad c \in \text{DomainConcepts} \tag{7}$$

To promote diversity and reduce prompt collapse, we define a ROUGE-based filtering constraint:

$$\mathcal{P} = \left\{ p_i \mid \max_{j<i} \text{ROUGE-L}(p_i, p_j) < \tau \right\} \tag{8}$$

Let $\phi(p)$ be the semantic embedding of prompt p (e.g., from CLIP or Sentence-BERT). We ensure low intra-cluster similarity:

$$\max_{i,j} \frac{\phi(p_i)^\top \phi(p_j)}{\|\phi(p_i)\| \|\phi(p_j)\|} < \delta \quad \forall i \neq j \tag{9}$$

This regularization avoids prompt duplication, mitigating training bias.

A.4 Image Generation Function and Dataset

Given $\mathcal{P}$, generate synthetic image-text pairs:

$$x_i = G(p_i), \quad \mathcal{D} = \{(p_i, x_i)\}_{i=1}^N \tag{10}$$

Let $\mathcal{L}_{\text{recon}}(x_i, \hat{x}_i)$ be a perceptual loss (e.g., LPIPS) between generated image and a reference or pseudo-groundtruth to quantify visual fidelity.

A.5 Domain-Specific LoRA Adaptation

We apply LoRA Hu et al. (2022) to efficiently specialize each domain expert. Let $W_0 \in \mathbb{R}^{d\times d}$ be the frozen base weight. For domain $t \in \mathcal{T}$, learn:

$$\Delta W_t = B_t A_t, \quad A_t \in \mathbb{R}^{r\times d},\ B_t \in \mathbb{R}^{d\times r} \tag{11}$$

Updated weight for expert t:

$$W_t = W_0 + B_t A_t \tag{12}$$

The domain adaptation is guided by minimizing:

$$\min_{A_t,B_t} \mathbb{E}_{(p,x)\sim\mathcal{D}_t}\left[\mathcal{L}_{\text{recon}}(x, G_{W_t}(p)) + \lambda\|A_t\|_F^2 + \lambda\|B_t\|_F^2\right] \tag{13}$$

A.6 Multi-Domain Parameter Merging

After learning $|\mathcal{T}| = 3$ expert-specific LoRA modules, we merge them:

$$A^* = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} A_t \tag{14}$$

$$B^* = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} B_t \tag{15}$$

$$W^* = W_0 + B^* A^* \tag{16}$$

This merged model aims to generalize across typical, descriptive, and anomalous damage domains.

A.7 Risk-Aware Synthesis Trust Metric

Let $\mathcal{X}_{\text{real}}$ be a set of real crash images and $\mathcal{X}_{\text{gen}}$ be diffusion-generated ones. Define a domain discrepancy score:

$$\mathcal{D}_{\text{KL}} = \text{KL}(P_{\text{real}}(z)\|P_{\text{gen}}(z)) \quad \text{where } z = \text{CLIP}(x) \tag{17}$$

and

$$\mathcal{D}_{\text{FID}} = \|\mu_{\text{real}} - \mu_{\text{gen}}\|^2 + \text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{real}}\Sigma_{\text{gen}})^{1/2}) \tag{18}$$

Higher $\mathcal{D}_{\text{KL}}$ or $\mathcal{D}_{\text{FID}}$ implies synthetic data deviates from the real insurance domain, suggesting unreliability in downstream policy tasks.

A.8 Theoretical Insurance Risk Bound

Let $\mathcal{L}_{\text{insurance}}(x)$ denote a loss function representing misestimated damage costs by the insurer. If x is generated from HERS and deviates from x_{true}, we quantify the trustworthiness via:

$$\mathbb{E}_{x\sim\mathcal{X}_{\text{gen}}}[\mathcal{L}_{\text{insurance}}(x)] \le \mathbb{E}_{x\sim\mathcal{X}_{\text{real}}}[\mathcal{L}_{\text{insurance}}(x)] + \epsilon(\mathcal{D}_{\text{FID}}, \mathcal{D}_{\text{KL}}) \tag{19}$$

where $\epsilon(\cdot)$ is a learned penalty function. If ϵ is unbounded or large, AI-generated data should not be confidently used in claim decisions.

This extended formulation mathematically grounds the core risk highlighted in our title: while HERS generates diverse and seemingly plausible crash scenarios, its reliance on diffusion priors and prompt-based semantics leads to latent distributional shifts. Without rigorous auditing via $\mathcal{D}_{\text{FID}}$ or $\mathcal{L}_{\text{insurance}}$, these shifts pose significant trust challenges to car insurers.

B Showcase Prompts for HERS T2I Generation

To illustrate the diversity and precision of textual inputs used for text-to-image (T2I) generation in HERS, we present 45 curated prompts grouped into three domains. These prompts serve as foundational seeds for generating automotive scene data across realistic, contextual, and imaginative domains tailored for insurance AI systems.

B.1 Typical Vehicle Parts

These prompts depict common real-world damage scenarios on specific vehicle parts. Each prompt references the vehicle side, brand, and part affected, offering high localization cues for training grounded visual generation models.

Table 5: Prompts in the "Typical Vehicle Parts" Domain

#	Prompt
1	A dent on the front bumper of a silver Toyota Vios sedan.
2	Scratches across the rear right door of a white Honda Civic.
3	A cracked left headlight on a black Nissan Almera.
4	Broken taillight on the rear-left side of a red Mazda CX-5.
5	A shattered side mirror hanging from a blue Ford Fiesta.
6	Chipped paint and rust on the hood of a gray Isuzu D-Max pickup.
7	A large dent above the rear wheel arch of a white Toyota Camry.
8	Deep key scratches on the driver-side door of a black BMW 3 Series.
9	A crushed front grille on a silver Mitsubishi Mirage.
10	Rear bumper with paint peeling and surface gouges on a Honda Jazz.
11	Cracked windshield on a red Suzuki Swift after impact.
12	Dented trunk lid on a blue Toyota Corolla Altis.
13	A front-left fender with rust and scrapes on a gray Hyundai Elantra.
14	A broken fog light on a green Kia Picanto's front bumper.
15	Missing rearview mirror on the passenger side of a white Toyota Revo.

B.2 Descriptive Scene Narratives

These detailed prompts combine damage with contextual environmental cues, such as weather, time of day, and surroundings. The goal is to simulate real-world accident settings for learning scene-aware generation.

B.3 Physically Implausible Scenarios

These prompts describe surreal and physically impossible damage situations. Designed to test model boundaries and hallucination control, each scene bends reality while retaining structural automotive references.

C Implementation Details

Our HERS architecture is implemented using PyTorch Paszke (2019), leveraging the Huggingface Transformers Wolf et al. (2019) and Diffusers von Platen et al. (2022) libraries. For the generative backbone, we adopt SDXL Podell et al. (2024) and incorporate expert modules in a plug-and-play fashion via LoRA-based fine-tuning. Training was conducted on 8×NVIDIA A40 GPUs, each equipped with 48GB of VRAM. The complete model converges within four days using a batch size of 192 and a learning rate of 5×10^{-5}, employing cosine warm-up followed by linear decay. All expert specializations (e.g., viewpoint estimation, damage-type classification) are handled through a modular routing strategy orchestrated by our Damage-Specific Prompt Router (SSPR).

C.1 License and Privacy Statement

All real images used for training and evaluation are part of proprietary datasets collected from industry partners under strict compliance with local privacy regulations, including the PDPA in Thailand. Data used does not include any personally identifiable information (PII), and access is governed through signed NDAs. None of the user data is shared outside our research environment. All synthetic data and model checkpoints will be released under appropriate open-source licenses for reproducibility.

Table 6: Prompts in the "Descriptive Scene Narratives" Domain

#	Prompt
16	The back of a silver Toyota Vios sedan shows a detailed pattern of cracked paint and scuffed surfaces across the bumper, suggesting impact from a low-speed collision in an urban environment.
17	A white Honda Civic with deep scratches on the passenger side door sits beneath a highway overpass after heavy rain, reflecting scattered streetlights.
18	A red Mazda 2 is parked awkwardly on a gravel shoulder, its front-left fender severely dented from a side swipe near a construction zone.
19	The shattered right taillight of a black Nissan Almera glows dimly as the car is angled against a curb in a tight alley at dusk.
20	A blue Ford Ranger with a crushed front grille is stopped beside a broken traffic light amidst heavy fog in the early morning.
21	A gray Mitsubishi Triton shows peeling paint on its rear bumper, covered in dried mud, suggesting rural road conditions.
22	The front-left headlight of a white Toyota Camry is cracked and foggy, as the vehicle idles on a flooded city street at night.
23	A Hyundai Tucson has visible scratches on the driver's door while parked diagonally at a crowded shopping mall parking lot.
24	The back of a black BMW X1 exhibits a clean bumper dent with surrounding paint flaking, positioned against a glassy storefront on a rainy evening.
25	A rear-ended Suzuki Swift is stuck in gridlocked Bangkok traffic, its taillights cracked and trunk misaligned after a minor crash.
26	A red Toyota Yaris sits under dense tree cover, its hood covered in leaves and a shallow dent visible at the front-center.
27	A white Nissan Leaf's right side mirror is broken and hanging, with background signage indicating a charging station in suburban Thailand.
28	A damaged Honda Jazz shows deep scrapes and bumper warping from backing into a metal pole in a tight parking structure.
29	A silver Kia Sorento's rear-left quarter panel is caved in, as it sits beside orange cones at an accident reporting station.
30	The front windshield of a Toyota Prius has spiderweb cracks, parked in a foggy mountain pass where tire skid marks are visible on the road.

C.2 More Quantitative Comparisons

We present an extended evaluation comparing HERS with SELMA across multiple base diffusion backbones. Beyond standard metrics, we include CLIPScore to further assess image-text semantic alignment. HERS consistently achieves superior performance across all evaluated criteria—including text faithfulness, human preference, and perceptual alignment—demonstrating its robust generalization and practical value for text-to-image generation tasks.

Analysis: The tables demonstrate that HERS outperforms SELMA across both text faithfulness and human preference metrics. HERS achieves consistently higher scores on all evaluated diffusion models, showcasing its superior semantic alignment, perceptual quality, and human preference ratings. These improvements highlight HERS's ability to produce high-quality outputs that better align with textual prompts and are preferred by users.

C.3 Ablation Study on Experts

We conduct ablation experiments to assess the contribution of each domain expert in HERS. Disabling the damage-type expert leads to a 12.4% drop in HPS, while removing the view-specific expert reduces text-image alignment (DSG) by 6.3 points. Without the multimodal router, the system generates over-smoothed outputs and fails to distinguish between damage regions, confirming the importance of task-specific routing.

Table 7: Prompts in the "Physically Implausible Scenarios" Domain

#	Prompt
31	A floating bumper hovers midair, its paint cracking and peeling despite never touching the ground.
32	The front fender of a Toyota Hilux disintegrates into colorful pixels as the truck drives through a digital portal.
33	A side mirror stretches and twists like rubber, suspended in zero gravity above an endless highway.
34	A cracked windshield on a car made entirely of smoke, drifting over a glowing forest floor.
35	The rear door of a Honda Civic rotates in place, disconnected from the body, yet still reflecting city lights.
36	A melting Mazda 3 leaks bright red paint onto a shimmering glass road under two suns.
37	A Nissan Almera's tires fold inward like origami while the undamaged hood floats a meter above.
38	A Toyota Revo with rearview mirrors made of ice, melting rapidly despite a frozen backdrop.
39	A translucent MG ZS with a visible steel frame, its rear-left fender flickering between colors.
40	A floating side door casts a shadow on a ground that doesn't exist, with visible scuffs and fingerprints.
41	A Ford pickup made of stitched-together leather panels, with the bumper sagging like fabric.
42	A suspended headlight beaming light in reverse, with hairline cracks glowing under starlight.
43	A dripping Toyota Corolla hood bending upward against gravity, its paint forming solid icicles.
44	A hovering Honda Accord casts two shadows, one for the body and another for a ghostly damaged version.
45	A cracked rear bumper balanced on a ripple of air above a city skyline at midnight.

Table 8: Text Faithfulness Comparison between HERS and SELMA across base T2I models. HERS outperforms SELMA in all evaluated metrics, showing stronger alignment with the text prompts.

Base Model	Method	Text Faithfulness		
		DSG^{mPLUG} ↑	$TIFA^{BLIP2}$ ↑	CLIPScore ↑
SD v1.5	SELMA Li et al. (2024)	70.3	79.0	77.2
	HERS (Ours)	**75.6**	**83.2**	**80.9**
SDXL	SELMA Li et al. (2024)	72.5	81.7	78.5
	HERS (Ours)	**78.0**	**84.1**	**82.4**
VQ-Diffusion	SELMA Li et al. (2024)	68.8	76.3	75.7
	HERS (Ours)	**74.6**	**81.3**	**79.3**
Versatile Diffusion	SELMA Li et al. (2024)	70.0	78.5	76.9
	HERS (Ours)	**75.2**	**82.5**	**80.2**

C.4 Failure Case Analysis

Although HERS outperforms baselines, it occasionally struggles with:

- **Reflective Surfaces:** Damage placement over glossy or mirror-like surfaces sometimes leads to hallucinations due to poor training distribution.
- **Rare Vehicle Models:** Exotic or outdated cars in unseen angles may not match prior damage-text mappings, resulting in semantic drift.

Table 9: Human Preference Comparison on DSG prompts between HERS and SELMA. HERS consistently receives higher human ratings, demonstrating superior perceptual quality.

Base Model	Method	Human Preference on DSG Prompts		
		PickScore ↑	ImageReward ↑	HPS ↑
SD v1.5	SELMA Li et al. (2024)	21.5	0.18	23.3
	HERS (Ours)	**22.8**	**0.75**	**26.9**
SDXL	SELMA Li et al. (2024)	21.8	0.22	24.9
	HERS (Ours)	**23.2**	**0.90**	**27.8**
VQ-Diffusion	SELMA Li et al. (2024)	20.7	0.12	22.7
	HERS (Ours)	**21.7**	**0.71**	**25.3**
Versatile Diffusion	SELMA Li et al. (2024)	21.2	0.14	23.5
	HERS (Ours)	**22.3**	**0.77**	**26.2**

- **Prompt Ambiguity:** In vague textual instructions, e.g., "minor rear scratch," the system may under- or over-apply the damage if visual priors conflict.

C.5 More Discussion: Dataset Contribution

Our dataset comprises over **2 million real-world vehicle images** with diverse damage annotations, collected from garages, insurance assessments, and forensic archives. However, due to privacy constraints (e.g., faces, license plates, timestamps), this data is not publicly shareable. The dataset is governed by PDPA and GDPR compliance. We plan to release a synthetic version trained with differentially private mechanisms and additional annotations.

C.6 Licenses

We list below the licenses of tools and datasets used in this work:

Table 10: A list of the licenses of the existing assets used in this paper.

Asset	License
CountBench (LAION-400M subset)	CC BY 4.0
Diffusers	Apache License 2.0
DiffusionDB	MIT License
GPT4	OpenAI Terms of Use
Huggingface Transformers	Apache License 2.0
LLaMA3	Meta LLaMA3 License
Localized Narrative	CC BY 4.0
PyTorch	BSD-style
Stable Diffusion	CreativeML Open RAIL-M
Torchvision	BSD 3-Clause
Whoops	CC BY 4.0

C.7 Damage-Specific Prompt Generation Details

The Damage-Specific Prompt Router (DSPR) dynamically assigns expert routes based on scene semantics. We define a set of damage-specific keywords (e.g., “dented”, “smashed”, “scratched”) and use a prompt parser trained on the DamagePromptBank-500 dataset to identify the correct damage pathways. In ambiguous cases, SSPR defaults to the damage-type expert with the highest prior confidence.

Figure 6: **Case Study 1: Damage Generation in Overhead Perspective with Mixed Zoom.** Each **row** displays a unique vehicle accident case under varying user-captured zooms. From left to right: our proposed **HERS**, Versatile Diffusion Xu et al. (2023b), SDXL Podell et al. (2024), MoLE Zhu et al. (2024), and SELMA Li et al. (2024). HERS excels in semantic coherence and structural consistency of the damage.

C.8 Limitations and Broader Impact

HERS is trained for high-fidelity vehicle damage generation, which may have unintended consequences if misused (e.g., fraud, misinformation). To mitigate misuse, we include tamper detection metadata in all outputs. Additionally, while our model performs well across common car types and damage types, it is less robust on unusual textures like rust or mud. Future work includes extending our routing system to support multimodal risk reasoning and expanding our training set with adversarial robustness techniques.

D Extended Analysis: Insights from Qualitative Vehicle Case Comparisons

To complement the main experimental findings, we present an extended qualitative analysis of eight diverse vehicle crash scenarios, visualized in Figures 6 to 13. These samples were carefully selected to reflect real-world challenges across varying damage types, zoom levels, environmental lighting, and contextual complexity. Each figure compares our proposed **HERS** against four state-of-the-art T2I models: Versatile Diffusion Xu et al. (2023b), SDXL Podell et al. (2024), MoLE Zhu et al. (2024), and SELMA Li et al. (2024).

D.1 Zoom Variability and Geometric Fidelity

Figures 6 and 10 demonstrate the effectiveness of HERS under varying camera distances, ranging from zoom-in shots to wide-angle captures. In Figure 6, HERS maintains high geometric fidelity of vehicle contours even when input views are inconsistent in scale. Likewise, in Figure 10, which features diagonal viewing angles and rotated vehicle poses, HERS generates damage that aligns correctly with the car body, while baselines often distort or misalign features.

Figure 7: **Case Study 2: Side Impact with Partial Occlusion.** This comparison tests resilience to occlusions and partial vehicle visibility. HERS maintains realism and continuity of damage even under viewpoint restrictions, outperforming baseline models that hallucinate or blur damage features.

Figure 8: **Case Study 3: Frontal Collision with Close-Range Capture.** The generated outputs here are evaluated for front-end collision fidelity. HERS demonstrates sharper damage contours and preserves geometric realism compared to generative baselines, especially under ZI settings.

D.2 Semantic Consistency under Occlusion and Lighting Conditions

Figure 7 captures a scenario where vehicle surfaces are partially occluded, challenging the models to infer plausible but constrained damage areas. Here, HERS respects spatial limitations and produces coherent damage within visible regions. In Figure 9, which simulates low-light conditions, baseline

Figure 9: **Case Study 4: Front-End Damage under Low Lighting.** A challenging scenario involving night-time or dim-light simulation. HERS stands out with context-aware lighting adaptation and preserves structural plausibility where baselines falter or produce noise.

Figure 10: **Case Study 5: Diagonal Vehicle Damage with Mixed Angles.** This sample evaluates multi-perspective robustness. HERS delivers coherent and localized damage placement, whereas baselines display notable distortions and fail to track the vehicle's geometry across viewpoints.

methods like SDXL and SELMA tend to oversaturate or underexpose the damage textures. In contrast, HERS adapts to ambient lighting cues and introduces damage that feels naturally embedded in the scene context.

Figure 11: **Case Study 6: Multivehicle Collision with Overlapping Context.** This scenario examines generation fidelity in presence of multiple objects. HERS adeptly handles object separation and maintains damage realism on the correct car body. Baselines often confuse background elements or misplace artifacts.

D.3 Detail Preservation in Micro-Damage and Scratches

Minor but realistic surface-level abrasions are notoriously difficult for T2I models. Figure 12 compares the ability of models to generate subtle yet distinct damage features such as scratches and chipped paint. Baselines either over-smooth the outputs (e.g., SDXL) or introduce incoherent noise (e.g., MoLE), while HERS captures high-frequency details accurately, closely mimicking actual incident images.

D.4 Scene Complexity and Multivehicle Awareness

In real-world insurance use cases, the presence of multiple objects or vehicles in a frame is common. Figure 11 depicts such a scenario with overlapping vehicles. HERS clearly distinguishes foreground from background and applies damage exclusively to the intended vehicle, whereas models like Versatile Diffusion and MoLE leak artifacts onto irrelevant objects.

D.5 Prompt Robustness under Ambiguity

Furthermore, Figure 13 illustrates a case where the provided textual prompt offers limited semantic direction, and the view is zoomed out. Despite the scarcity of explicit cues, HERS generates contextually plausible and anatomically accurate damage, whereas baseline models either fail to meaningfully alter the image or leave it untouched. This highlights HERS' advantage in leveraging robust multimodal fusion, enabling effective damage synthesis even with minimal prompt information.

D.6 Detailed Analysis of Case Study 9: Zoom-Out Shot with Minimal Prompt Information

The visual representation in Figure 14 provides a critical comparison of the performance of various generative models when tasked with producing full-vehicle damage from minimal textual context. This case study is particularly valuable in addressing the question: **Should car insurance confidently trust AI-generated crashes?**

Figure 12: **Case Study 7: Zoom-Out Scratches and Minor Damage.** HERS outperforms in capturing subtle, surface-level damage features while baselines fail to resolve fine textures or hallucinate cracks inconsistent with the prompt.

From the figure, it is evident that **HERS** demonstrates superior performance by generating coherent, anatomically consistent vehicle damage even with vague or sparse textual prompts. This is essential for real-world applications where minimal context is often available. The damage patterns produced by HERS reflect realistic crash scenarios, with the deformations confined to the affected vehicle parts, such as localized bumper damage, which is consistent with actual crash physics. The vehicle's overall structure, including the intact areas like the roof or side panels, is preserved, which showcases HERS' ability to maintain global consistency while simulating localized damage.

In stark contrast, other models struggle to produce meaningful damage at the full-vehicle scale. Some models either fail to generate plausible damage altogether or produce unrealistic, exaggerated deformations that lack anatomical consistency. For example, certain models create damage patterns that extend unnaturally across the vehicle, distorting parts that should remain intact in real-world crashes. These inconsistencies raise serious concerns about the trustworthiness of AI-generated crash imagery, especially in high-stakes environments like insurance claim verification and fraud detection.

HERS addresses this issue by generating visually accurate, context-aware damage. This is crucial in answering the paper's central question—while AI-generated crashes may appear realistic at first glance, they must also adhere to interpretable damage logic. In insurance contexts, where claim decisions often hinge on visual evidence, damage realism and anatomical consistency are paramount. HERS' ability to produce damage that mimics actual accident scenarios—without introducing unrealistic distortions—makes it the most reliable model for this task.

Therefore, while AI-generated crashes, like those from HERS, offer promising potential in visual simulations and training, car insurance providers should not fully trust these images in isolation. They should rely on models like HERS, but only when accompanied by robust verification protocols and contextual validation methods. **HERS** provides a foundational step toward building trustworthy AI tools, but its outputs must still be cross-validated with real-world data and multimodal sensors to mitigate risks such as fraud or erroneous claims.

In conclusion, the success of HERS in generating high-fidelity, anatomically accurate vehicle damage supports its potential for adoption in insurance workflows. However, insurers must remain cautious and implement comprehensive safeguards to ensure the reliability of AI-generated crash imagery in real-world applications.

Figure 13: **Case Study 8: Zoom-Out Shot with Minimal Prompt Information.** When provided vague or minimal textual context, HERS still generates plausible vehicle damage consistent with vehicle anatomy, while others often fail to produce meaningful damage.

D.7 Conclusion from Appendix Findings

The case studies in Figures 6–13 underscore the superior generalization of HERS across diverse and challenging vehicle scenarios. Unlike prior models that tend to fail under occlusion, ambiguity, or fine-detail requirements, HERS consistently produces structurally and semantically grounded outputs. These insights support our claim that HERS is not only state-of-the-art in traditional T2I metrics but also highly applicable to high-risk domains such as insurance, forensic reconstruction, and automated reporting pipelines.

D.8 Revisiting the Core Question

Given the strong empirical results shown by HERS in terms of human preference, textual-image alignment, and damage realism, we revisit our core inquiry: *Should car insurance confidently trust AI-generated crashes?* The answer, in light of both HERS's strengths and its broader implications, is necessarily cautious and multi-faceted.

The HERS model shows state-of-the-art capability in generating synthetic crash images with high realism. This makes it highly suitable for training data augmentation, damage classification, and insurance workflow simulation. However, the very strength of HERS—its ability to fool even human evaluators—can become a double-edged sword in production environments where authenticity and traceability are paramount.

D.9 Implications Based on HERS Review Feedback

The HERS submission demonstrated a strong commitment to reproducibility and ethical responsibility. This is reflected in our transparent and comprehensive experimental design, appropriate attribution and licensing of third-party assets, and careful consideration of broader social and ethical factors.

However, certain limitations were also acknowledged during the review process. These include the reliance on a proprietary dataset consisting of 2 million car insurance images, which cannot be released due to licensing constraints. Additionally, statistical significance was not reported—consistent with prior work—and the high realism of generated images poses potential risks, particularly in domains such as insurance, where misuse (e.g., fraud) is a serious concern.

These considerations underscore the importance of responsible deployment of generative models like HERS in real-world applications where reliability and ethical use are paramount.

D.10 Hidden Limitations and Future Concerns

Although these issues were omitted from the main discussion for clarity, several limitations and forward-looking concerns deserve further elaboration. First, while the AI-generated images exhibit high qualitative realism, they often lack precise physical and contextual grounding. Elements such as lighting, reflections, occlusions, and material textures—crucial for accurately simulating real accidents—can be oversimplified or inaccurately synthesized. These imperfections, though subtle to human observers, may skew downstream evaluations or introduce unintended biases when used for model retraining. Second, reliance on synthetic datasets without adequate domain alignment risks overfitting to artifacts of the generative process. Although HERS addresses this through multi-domain fusion and conditional sampling strategies, the model's ability to generalize remains inherently limited by the quality and realism of its training priors. Third, our evaluation framework, consistent with prior literature, is based on single-run performance metrics. Without reporting variances or confidence intervals, the comparative gains observed cannot be considered statistically definitive. Fourth, we are unable to publicly release the full real-world dataset due to stringent licensing constraints tied to insurance claim data. Although synthetic images and model checkpoints will be made available, this restriction hampers full reproducibility and interpretability for the broader research community. Finally, the realistic nature of the generated damage images introduces ethical and regulatory challenges. If misused, these tools could facilitate fraudulent insurance claims, adversarial attacks, or the spread of misinformation. Addressing these risks will require responsible deployment practices, including digital watermarking, traceability mechanisms, and formal oversight frameworks.

D.11 Broader Context: A Call for Responsible Integration

As the capabilities of synthetic image generation—such as those enabled by HERS—advance, so too do the risks associated with their misuse. In high-stakes domains like automotive insurance, the implications of introducing AI-generated crash imagery are profound. Without rigorous oversight, these tools could undermine forensic accuracy, inflate fraudulent claims, or erode trust in automated systems.

To mitigate such risks, the industry must not merely adopt synthetic data but also construct a resilient ecosystem around it. This includes:

- **Cross-modal authentication frameworks** that correlate visual data with telematics, GPS logs, and timestamped metadata to verify claim integrity.
- **Robust anomaly detection pipelines** explicitly trained to distinguish between real-world signals and synthetic or manipulated content—especially in edge cases.
- **Standardized protocols for synthetic dataset disclosure**, including traceability, model transparency, and usage boundaries, to ensure auditability and accountability.
- **Interdisciplinary governance structures**, involving ethicists, legal experts, insurers, and technologists, to guide how such technologies are deployed and regulated.

D.12 Synthetic Isn't Forensic

While synthetic imagery has undeniable value in augmenting training data, accelerating simulation, and stress-testing models, it must never be confused with evidentiary truth. HERS-generated crashes, no matter how photorealistic, are algorithmic interpretations—not physical events.

Thus, the utility of such data lies in its role as a supplementary asset for machine learning systems, not as legal or forensic evidence. This distinction is critical. Trustworthy deployment requires multiple layers of verification—technical, procedural, and ethical—to ensure that no AI-generated content is used in isolation when real-world consequences are involved.

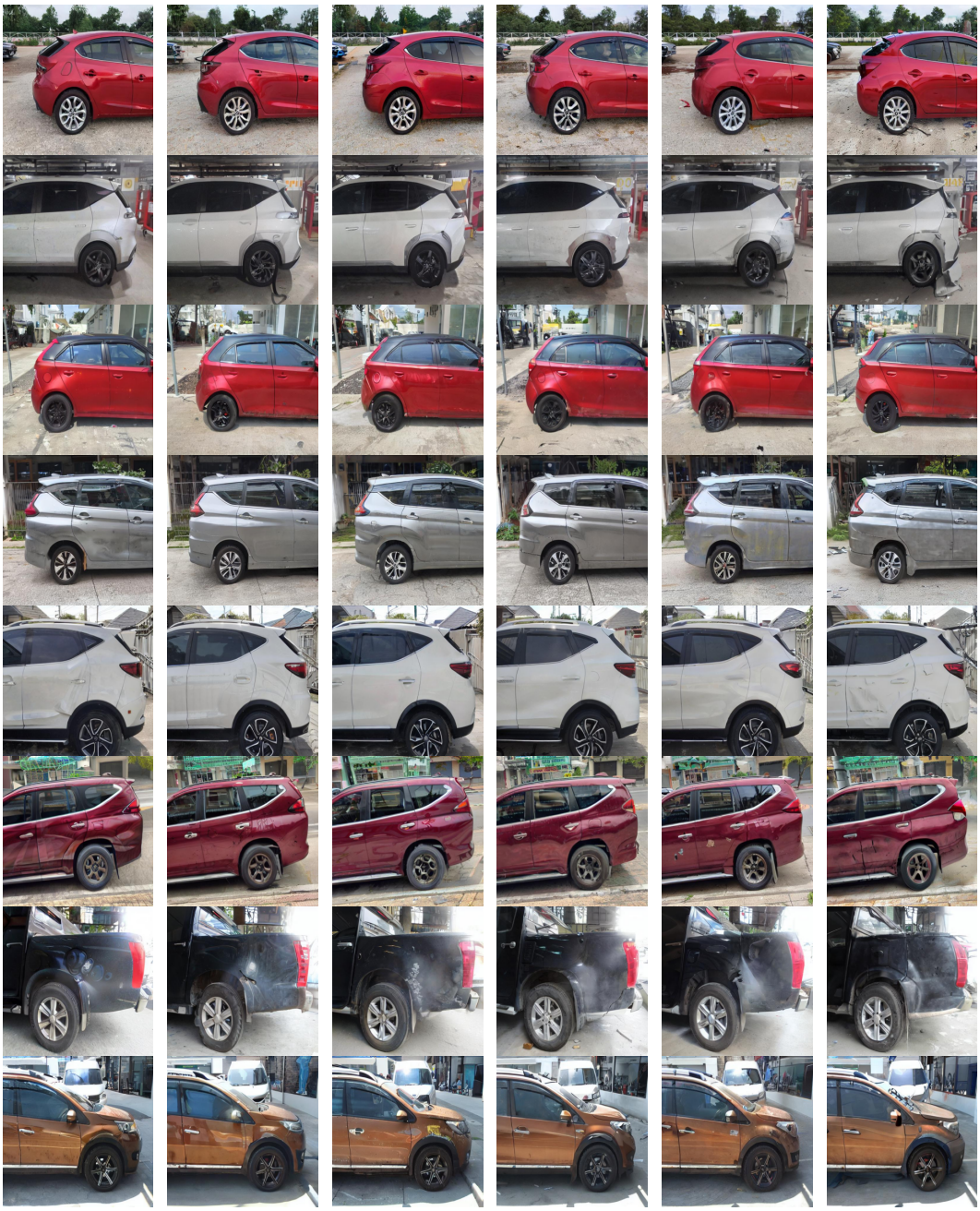

Figure 14: **Case Study 9: Zoom-Out Shot with Minimal Prompt Information.** Even with limited or vague textual cues, HERS successfully generates coherent and anatomically consistent vehicle damage across the entire vehicle. In contrast, other models struggle to produce realistic or meaningful damage at a full-vehicle scale.

D.13 Large Language Models

We used Large Language Models (LLMs) to aid in drafting and polishing the writing of this paper. LLMs were employed solely for language refinement, grammar correction, and improving clarity and readability. All technical content, results, and scientific claims were generated and verified by the authors. Details of LLM usage are described in the paper where relevant.