University Hassan I
National School of Applied Sciences of Berrechid (ENSAB)

Literature Review on Ethical Image Generation

Prepared by: Fatima Ezzahra Kninech
Khadija Mourad

Program: Information Systems Engineering and Big Data (ISIBD-S9)
Module: Data Mining
Supervised by: Prof. Hamid Hrimech

January 11, 2026

Abstract

Recent progress in diffusion-based image generation has greatly expanded access to visual content creation, enabling broader participation in creative domains. Yet this democratization raises pressing ethical challenges, including the risk of producing harmful material, reinforcing social biases, and violating established safety norms. Current safeguards—such as keyword filtering or human moderation—remain limited, offering neither fine-grained nor scalable solutions for multimodal regulation. Existing approaches that combine latent diffusion with toxicity classifiers often fail to capture the semantic nuances of user prompts or to ensure consistent ethical alignment between text and imagery. Moreover, opaque safety mechanisms tend to oscillate between excessive restriction and permissiveness, allowing problematic outputs to slip through. To address these gaps, this project introduces a modular pipeline built on a safety-by-design paradigm. By integrating ethical text analysis, domain classification, semantic enrichment, controlled generation, and visual validation, the system aims to deliver responsible image creation that balances technical performance with ethical robustness.

Keywords : Image Generation, Artificial Intelligence, Ethics, Deep Learning, Multi-Model Systems.

Contents

1 Introduction

Since the introduction of latent diffusion models in 2022, AI-driven image generation has reached an unprecedented level of realism. Trained on billions of text–image pairs, these models can translate abstract concepts into complex visual representations, reshaping industries such as design, marketing, and real estate.

This remarkable capability, however, comes with significant risks. Diffusion models can be misused to produce violent, hateful, or biased content. Ethical considerations are therefore no longer peripheral but a fundamental requirement for the professional deployment of such technologies.

Addressing these challenges demands more than a simple generation interface. It requires architectures that embed moderation directly into the computational pipeline. This involves a synergy between natural language processing (NLP), to interpret human intent, and computer vision, to evaluate the generated output.

This review explores the theoretical foundations of generative models and the mechanisms of ethical control, highlighting how assembling specialized models can outperform a single large-scale system in terms of safety, governance, and business relevance.

Main contributions of this work are summarized as follows:

1. The design of a semantic text-based filtering mechanism for proactive ethical control.
2. An automatic domain classification strategy applied jointly to textual prompts and generated images.
3. An automated prompt enrichment process to enhance semantic consistency and visual fidelity.
4. A domain-aware specialization mechanism for controlled visual generation.
5. A dual ethical validation framework operating at both textual input and visual output levels.

2 Theoretical Foundations of Image Generation

2.1 Principles of Deep Learning for Generation

Generative AI relies on the ability of a model to learn a complex data distribution in order to create new instances.

Supervised vs. Unsupervised Learning: While classification tasks (e.g., ethics, domain) depend on supervised learning with labeled data, image generation models leverage self-supervised learning. They learn to reconstruct data from Gaussian noise using image–text pairs as semantic guidance.

Latent Representations: A key concept is the latent space. Instead of working directly on pixels, modern models compress the image into a low-dimensional mathematical space. Within this space, the model manipulates concepts (e.g., “red dress,” “modern house”) before decoding them back into high-resolution images.

2.2 Image Generation Models

Four main architectures dominate the literature and structure this project:

- **GAN (Generative Adversarial Networks):** Historically the first successful models, based on competition between a generator and a discriminator. Although fast, they suffer from training instability and limited diversity compared to diffusion models.

- **Diffusion Models (e.g., Stable Diffusion):** The current state of the art. These models learn to reverse a degradation process (noise addition). Starting from pure noise, the model progressively “cleans” the image step by step to reveal coherent shapes guided by the prompt. The use of LoRA here serves as an optimization technique, allowing fine-tuned adjustments of diffusion model weights without retraining the entire network.

- **Vision Transformers (ViT):** Inspired by the success of Transformers in text, ViTs divide an image into patches (small squares) treated like words. This approach enables a global understanding of the image, which is crucial for the `image_ethics` model to detect anomalies or inappropriate content.

- **Text-to-Image Models:** These systems employ encoders (often CLIP or DistilBERT in this case) to align text and image within the same semantic space. The generation model then uses this textual “compass” to navigate the latent space and produce visuals corresponding to the description.

3 Ethical Challenges in Image Generation

AI-driven image generation raises major ethical, social, and technical challenges. Scientific literature consistently identifies several categories of risks that must be considered when developing and deploying large-scale generative models.

3.1 Algorithmic Bias

Diffusion models inherit biases from their training datasets, often composed of massive web-scraped collections such as LAION-5B. Numerous studies have shown that, without corrective mechanisms, these models tend to associate professions, social roles, or physical attributes with specific genders, ethnicities, or cultures. Such algorithmic biases reproduce and amplify existing societal stereotypes, posing direct risks to fairness and inclusivity.

3.2 Social Representation and Diversity

Ensuring fair representation is a central issue in generative AI ethics. Underrepresentation or biased portrayal of minority groups can lead to distorted or discriminatory visual outcomes. A responsible image generation system must be able to produce coherent visual diversity that respects cultural and social identities, avoiding caricature or symbolic exaggeration.

3.3 Sensitive and Inappropriate Content

Generative models can produce explicit violence, non-consensual nudity, or hateful imagery—risks frequently highlighted in the literature. Such NSFW (Not Safe For Work) content may be generated intentionally or emerge indirectly from ambiguous prompts. Recent research emphasizes proactive prevention strategies integrated directly into the generation pipeline.

3.4 Intellectual Property and Copyright

Automated image generation also raises complex intellectual property concerns. Training datasets containing copyrighted works risk implicitly reproducing artistic styles, particularly those of living artists. In professional contexts such as fashion or real estate, these issues demand heightened vigilance to ensure originality and legal compliance of generated outputs.

3.5 Security, Misuse, and Disinformation

Image generation can be misused for malicious purposes, including the creation of deepfakes or falsified visual evidence. Such abuses threaten public trust and informational stability. Literature highlights the need for control mechanisms capable of blocking the generation of public figures or scenes likely to contribute to disinformation.

4 Existing Methods for Ethical Image Generation

In response to these risks, research proposes strategies to embed ethical principles into generative systems. These approaches can be structured across three complementary levels: data, model training, and external control mechanisms.

4.1 Data Preparation and Filtering

The first level of ethical intervention occurs during dataset construction. Large-scale diffusion training corpora undergo curation processes to remove illegal, toxic, or low-quality content. Given the scale of these datasets, manual filtering is impractical, leading to reliance on automated scoring and classification methods. Ethical annotation of data, performed via automated classifiers, categorizes images and text descriptions according to safety compliance. Additionally, class-balancing techniques such as targeted data augmentation are employed to mitigate demographic imbalances and reduce statistical bias.

4.2 Ethical Training Methods

Beyond data, some approaches integrate ethical constraints directly into model training. Controlled fine-tuning, particularly through Low-Rank Adaptation (LoRA), specializes pre-trained models on validated subsets, reducing exposure to problematic content. Other methods introduce regularization mechanisms into the loss function to minimize performance disparities across demographic groups, enhancing overall fairness. Multi-task learning is also explored, where models are trained simultaneously to generate images and estimate safety or toxicity scores, enabling dynamic adjustment of the generative process. Transfer learning from robust pre-trained models further improves semantic understanding and reduces misinterpretation risks in domain-specific adaptation.

4.3 Auxiliary Ethical Control Models (Guardrails)

A third category of approaches relies on auxiliary models acting as supervisory guardrails around the main generator. These typically include textual and visual classifiers that analyze user prompts and generated outputs. In cases of non-compliance, such systems can block, modify, or reject results. Some architectures employ multimodal models to measure semantic proximity between prompts and sensitive concepts, enabling indirect detection of problematic intentions. Finally, post-generation safety models apply corrective mechanisms such as blurring or removal, ensuring that the final content delivered to users adheres to established ethical standards.

5 Project Architecture: Multimodal and Modular Pipeline

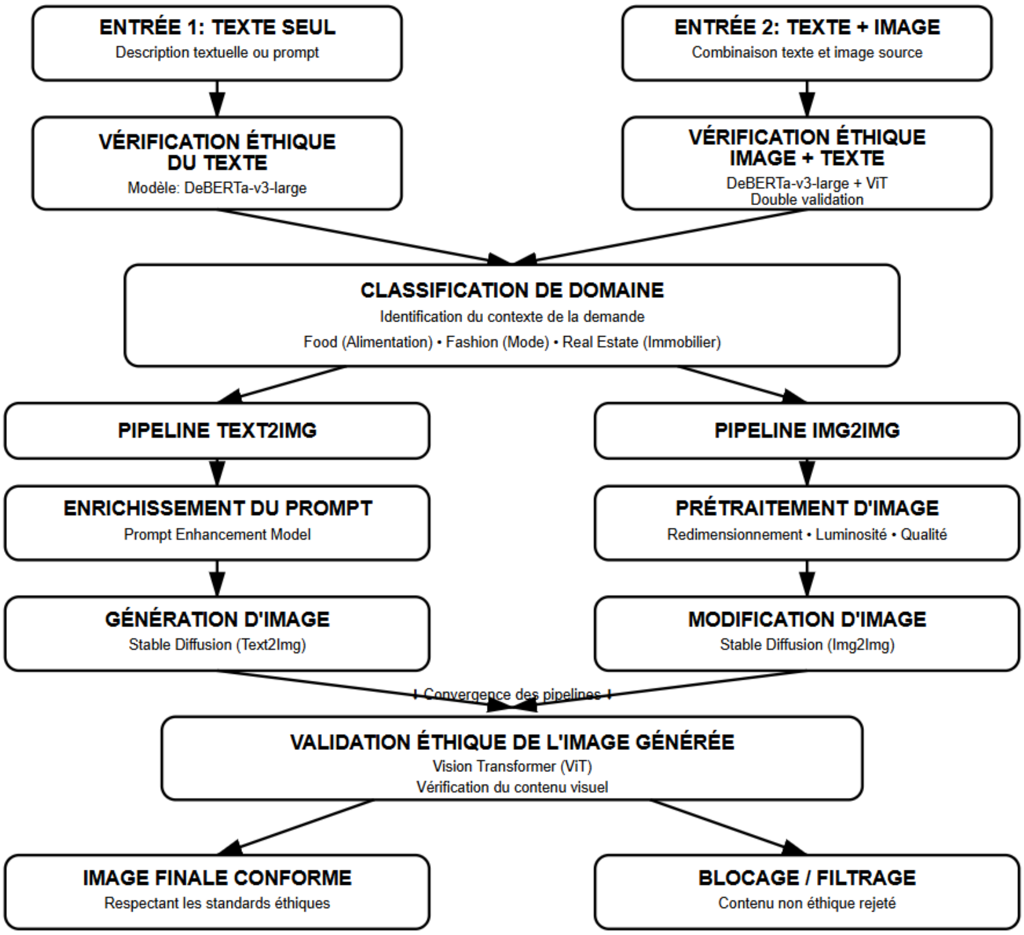
ENTRÉE 1: TEXTE SEUL
Description textuelle ou prompt
ENTRÉE 2: TEXTE + IMAGE
Combinaison texte et image source
VÉRIFICATION ÉTHIQUE DU TEXTE
Modèle: DeBERTa-v3-large
VÉRIFICATION ÉTHIQUE IMAGE + TEXTE
DeBERTa-v3-large + ViT
Double validation
CLASSIFICATION DE DOMAINE
Identification du contexte de la demande
Food (Alimentation) • Fashion (Mode) • Real Estate (Immobilier)
PIPELINE TEXT2IMG
PIPELINE IMG2IMG
ENRICHISSEMENT DU PROMPT
Prompt Enhancement Model
PRÉTRAITEMENT D'IMAGE
Redimensionnement • Luminosité • Qualité
GÉNÉRATION D'IMAGE
Stable Diffusion (Text2Img)
MODIFICATION D'IMAGE
Stable Diffusion (Img2Img)
Convergence des pipelines
VALIDATION ÉTHIQUE DE L'IMAGE GÉNÉRÉE
Vision Transformer (ViT)
Vérification du contenu visuel
IMAGE FINALE CONFORME
Respectant les standards éthiques
BLOCAGE / FILTRAGE
Contenu non éthique rejeté

Figure 1: Diagram of the multimodal and modular pipeline for ethical image generation.

The proposed architecture is built on a *Safety-by-Design* principle, orchestrating multiple specialized models to ensure both ethical integrity and visual consistency in generated and edited images. Unlike monolithic systems, this pipeline is divided into three main phases: intent validation (Input), generation and editing, and post-generation validation (Output).

5.1 Intent Validation (Input Layer)

Each user request is first analyzed by a text classifier based on DeBERTa-v3-large, capable of detecting non-compliant content and mitigating risks of textual filter bypassing (*jailbreaking*). At the same time, a domain classifier routes the request to one of the targeted business sectors (e.g., Food, Fashion, Real Estate), enabling the application of compliance rules and domain-specific visual styles.

5.2 Enhanced Generation and Editing

To improve semantic and visual fidelity, a *Prompt Enhancement* module based on the T5 architecture transforms concise instructions into enriched prompts. These prompts are then passed to the diffusion model for image generation or editing.

- **Text-to-Image (Text2Image)**: The enriched prompt guides the generation process to produce a complete image.
- **Image-to-Image (Img2Img)**: When a source image is provided, it undergoes preprocessing (resizing, brightness adjustment, and sharpening) before being merged with the enriched prompt, ensuring visual coherence between the original image and the requested modifications.

5.3 Post-Generation Validation (Output Guardrails)

All images, whether generated or imported, are subjected to a final ethical analysis by a Vision Transformer (ViT-base-patch16). This model processes the image through semantic patches, enabling the detection of inappropriate or non-compliant content that may have bypassed initial textual filters.

5.4 Blocking and Feedback Mechanism

The pipeline operates as a closed loop: any detection of non-compliance, at either input or output, triggers an immediate blocking mechanism. This modular structure transforms the system into a trusted ecosystem, where visual quality and ethical compliance are tightly interconnected. Furthermore, the modular design allows future extensions of the pipeline through the integration of new specialized models or additional business domains.

6 Methodology

This section details the design and implementation of the ethical image generation pipeline. The approach relies on a modular orchestration of six deep learning models, designed to secure each stage of the data flow.

6.1 General Overview of the Pipeline

The pipeline adopts an “ethical sandwich” architecture, where image generation is framed by textual safeguards (upstream) and visual safeguards (downstream). The system is capable of handling two distinct flows depending on the type of input provided by the user,

The operational flow is divided into the following steps:

1. **Input and Intent Validation (Input Layer)**
   The system accepts two input configurations:

   - **Flow A (Text only):** The user submits a descriptive prompt.
   - **Flow B (Text + Image):** The user submits a source image accompanied by modification instructions.

   At the input stage, two levels of verification are activated:

   - **Textual Analysis:** The DeBERTa-v3-large model evaluates the ethical compliance of the prompt, while DistilBERT identifies the application domain (Fashion, Food, Real Estate).
   - **Source Image Analysis (if provided):** In Flow B, the uploaded image is immediately processed by the ViT classifier to ensure it does not contain unethical content before further processing.

2. **Semantic Optimization and Preprocessing**
   Once the intent is validated:

   - **Prompt Enhancement:** For textual requests, the T5 model enriches the initial prompt by adding professional terminology and domain-specific stylistic details.
   - **Visual Normalization (Flow B):** The source image undergoes technical preprocessing (optimal resizing, brightness correction, sharpening) to maximize compatibility with the diffusion engine.

3. **Specialized Generative Core**
   Processing is performed by the latent diffusion model Stable Diffusion v1.5, optimized with LoRA adapters:

   - **Text-to-Image Mode:** Generation of a complete image from noise, guided by the enriched prompt.
   - **Img2Img Mode:** Transformation of the source image by injecting new semantic elements while preserving the original spatial structure (with calibrated transformation strength).

4. **Output Validation and Feedback (Output Layer)**
   Before final display, the resulting image undergoes a final series of tests using Vision Transformer (ViT) models:

   - **Final Ethical Verification:** Detection of any non-compliant visual anomalies produced during generation.

- **Domain Consistency:** Validation that the image correctly belongs to the business sector identified earlier.

6.2 Data Preparation and Dataset Engineering

Model robustness relies on a rigorous data curation strategy:

- **Multi-source Datastreams:** For textual ethics, a hybrid dataset was built from NVIDIA Aegis, OpenAI Moderation Dataset, and Hateful Memes, totaling 20,978 entries.
- **Cleaning and Balancing:** An undersampling technique was applied to perfectly balance classes (9,232 examples per class), avoiding prediction bias toward the majority class.
- **Vision Data Curation:** For visual validation, specialized datasets were merged: FairFace (ethics), VLSBench (illegal activities/erotica), and custom datasets from Kaggle and GitHub for violent content.
- **Domain Data:** A dataset of 38,998 images was segmented into three key domains (Fashion, Food, Real Estate) to train domain classifiers with maximum accuracy.

6.3 Model Selection and Justification

The choice of architectures was guided by a trade-off between semantic performance and computational efficiency:

Module	**Model**	**Justification**
Text Ethics	DeBERTa-v3-large	Selected for its superior ability to capture context and linguistic nuances compared to BERT, crucial for detecting "jailbreaking".
Text Domain	DistilBERT	Provides minimal latency for fast domain classification (100% accuracy in validation).
Prompt Enhancement	T5-small	Encoder-Decoder architecture ideal for text transformation, enabling the addition of professional terminology.
Generation	SD v1.5 + LoRA	Stable Diffusion v1.5 ensures stable generation; LoRA (Rank 4, Alpha 8) allows specialization without the cost of full retraining.

Table 1: Modules, Models, and Justifications

Module	Model	Justification
Image Validation	ViT-base-patch16	The Vision Transformer analyzes images by patches, offering finer detection of unethical details than traditional CNNs.

Table 1: Modules, Models, and Justifications

6.4 Model Training

Fine-tuning and LoRA Optimization: The generation model was trained using LoRA (Low-Rank Adaptation). With only 500 steps at a learning rate of $1e^{-4}$ on the Stable Diffusion UNet, domain-specific knowledge (Food, Fashion, Real Estate) was injected. This method achieved professional visual quality while maintaining a reduced GPU memory footprint.

Ethical and Domain Classifiers: Text models were trained for 6 epochs with a learning rate of $1e^{-5}$. Early Stopping was applied to halt training upon F1 metric convergence, ensuring robustness against overfitting. Vision Transformers were trained at a resolution of 224×224. The image domain classifier achieved an overall accuracy of 99.89%, demonstrating near-perfect separability of the chosen domains.

6.5 Pipeline Integration and Feedback Management

Software integration (via FastAPI) ensures synchronous communication between models. The system enforces a strict blocking policy: if the probability score for the "non_ethical" class exceeds a predefined threshold (Softmax > 0.5), the process is immediately interrupted. In Img2Img mode, specific preprocessing (resizing to multiples of 8, RGB conversion) is applied to ensure compatibility with the diffusion pipeline, while subjecting the source image to the same ViT validation as the generated output.

6.6 Evaluation and Performance Metrics

System effectiveness is measured through several key indicators:

- **Ethical Robustness:** Evaluated via confusion matrices showing an extremely low false negative rate (only 5 errors out of 3,225 ethical cases for the ViT model).
- **Semantic Fidelity:** Measured by ROUGE score (8.93%) for the prompt generator and CLIPScore (29.20) for text-image correlation, ensuring that the produced image faithfully matches the user request.

- **Domain Accuracy:** Validated by a classification report showing an F1-score of 1.00 across the three target domains, confirming successful specialization of the pipeline.

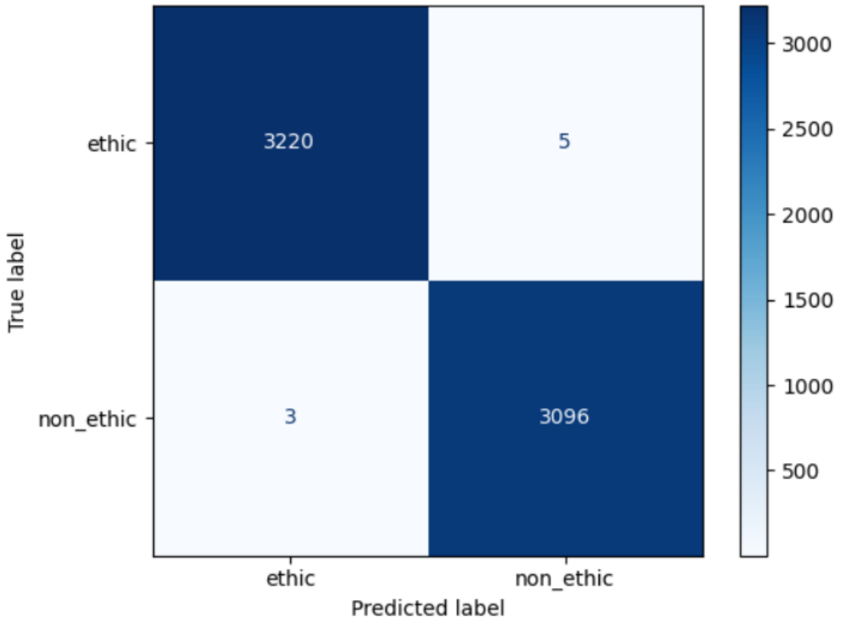

Figure 2: Confusion matrix of the image ethics classification model (ViT-base-patch16).

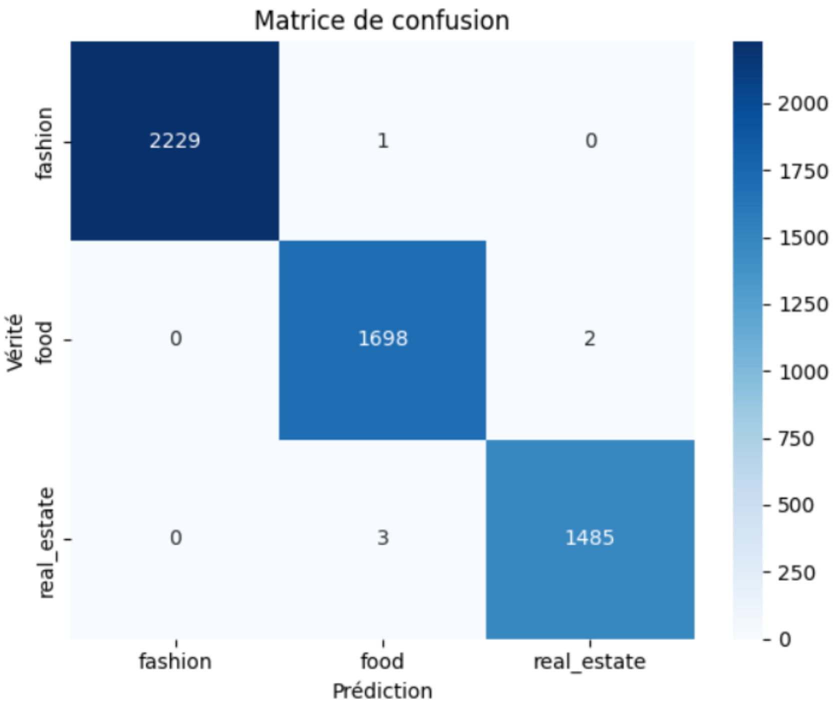

Figure 3: Confusion matrix of the image domain classification model (ViT).

7 Comparison with Existing Approaches

7.1 Scientific Positioning

Most current image generation systems (such as vanilla Stable Diffusion or DALL-E) rely on external safety filters that are often opaque or based on lists of forbidden words. This project positions itself as an integrated trust architecture, where moderation is not a peripheral option but a structural component of the computational pipeline.

7.2 Comparative Table of Approaches

Characteristics	**Standard Models (DALL-E, SD)**	**Research Approaches (API Filters)**	**This Project (Multi-model Pipeline)**
Architecture	Single massive model	Model + External Filter	Multi-model orchestration (6 models)
Ethical Control	Often limited to keywords	Post-filtering only	Double barrier (Text + Image)
Specialization	Generalist (sometimes imprecise)	Weak	Business domains (Fashion, Food, Real Estate)
Optimization	Raw query	Manual (User-side)	Automated Prompt Enhancement (T5)
Interface	Often API-only	Technical scripts	Full-Stack Interface (React/FastAPI)
Img2Img Mode	Available but poorly regulated	Rarely secured	Ethical verification of source image

Table 2: Comparison of Image Generation Approaches

8 Discussion and Limitations

8.1 Critical Analysis

Although the system demonstrates high performance (Accuracy $>$ 99% on certain modules), several technical and ethical challenges remain:

- **Computational Cost:** Running six models in parallel (DeBERTa, DistilBERT, T5, SD, and two ViTs) requires strict GPU resource management. In a production environment, sequential inference may introduce latency, although the use of distilled models (DistilBERT, T5-small) helps mitigate this risk.

- **Domain Scalability:** Currently limited to three domains, adding new sectors (e.g., Medical, Automotive) requires a data curation phase and additional LoRA fine-tuning, making system expansion dependent on the availability of specialized datasets.
- **Remaining Ethical Limits:** No automated system is infallible. Complex semantic jailbreaking or ambiguous images (irony, metaphor) remain challenges for current classification models, which handle ethics in a probabilistic manner.

9 Conclusion of the Review

9.1 Final Synthesis

This literature review has traced the evolution of image generation, moving from a pursuit of pure visual realism to a necessity for control and safety. Existing work highlights that the technical performance of a diffusion model cannot be dissociated from its social responsibility.

9.2 Contributions of the Proposed Methodology

The presented project provides a concrete response to the limitations of monolithic models through:

- Implementation of an *Ethical Sandwich* that secures data from input (text) to output (image).
- Use of LoRA techniques for efficient and energy-saving business specialization.
- Cross-validation of Domain/Ethics via Vision Transformer (ViT) architectures, ensuring consistency between user intent and the generated visual output.

9.3 Future Directions

To go further, the pipeline could be enriched by integrating automatic *Negative Prompting*, where the ethical classifier itself injects safety constraints into the diffusion model. Similarly, adding a watermark detection function on generated images would ensure traceability of synthetic content, thus addressing future international AI regulations (such as the European AI Act).

10 Bibliography

10.1 Natural Language Processing Architectures (NLP and Text Ethics)

- He, P., Gao, J., & Chen, W. (2021). *DeBERTaV3: Improving DeBERTa Using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding.* arXiv preprint arXiv:2111.09543.

- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT: A Distilled Version of BERT — Smaller, Faster, Cheaper and Lighter.* arXiv preprint arXiv:1910.01108.

- Raffel, C., Shazeer, N., Roberts, A., et al. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research, 21(140), 1–67.

- Kiela, D., Firooz, H., et al. (2020). *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes.* Advances in Neural Information Processing Systems (NeurIPS).

10.2 Image Generation and Diffusion Models

- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.

- Hu, E. J., Shen, Y., Wallis, P., et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models.* arXiv preprint arXiv:2106.09685.

- Zhang, L., Rao, A., & Agrawala, M. (2023). *Adding Conditional Control to Text-to-Image Diffusion Models.* Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

10.3 Computer Vision and Image Classification

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* International Conference on Learning Representations (ICLR).

- Kärkkäinen, K., & Joo, J. (2021). *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation.* IEEE Winter Conference on Applications of Computer Vision (WACV).

10.4 AI Ethics, Bias, and Content Safety

- NVIDIA. (2023). *Aegis: A Content Safety Dataset for Large Language Models.* NVIDIA Technical Report / Hugging Face Repository.

- Bianchi, F., Kalluri, T., et al. (2023). *Easily Fooled into Revealing Biases: Visual Adversarial Attacks on AI Safeguards.* arXiv preprint arXiv:2308.12303.

- WalledAI. (2024). *OpenAI Moderation Dataset: A Benchmark for Multimodal Safety.* Hugging Face Open-Source Dataset.

10.5 Frameworks and Engineering Tools

- Tiangolo, S. (2018). *FastAPI: High Performance, Easy to Learn, Fast to Code, Ready for Production.* Available online: `https://fastapi.tiangolo.com`.

- Wolf, T., et al. (2020). *Transformers: State-of-the-Art Natural Language Processing.* Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.