Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.32588

Joint Multimodal Data Desensitization Mechanism Based on Face-Swapping and Cross-Modal Semantic Alignment

Yanfeng Lu[1*], Shuchang Luo[1], Mingxin Chen[1], Pengxu Fang[1], Yusong Gao[1]

[1]School of Mathematical Sciences, Ocean University of China, Qingdao, China
**Corresponding Author. Email: luyanfeng@stu.ouc.edu.cn*

Abstract. With the explosive growth of multimedia data, the preservation of privacy in multimodal data has become a key research challenge. The traditional desensitization approach for a single data modality often fails to consider hidden privacy correlations between multiple data modalities, resulting in either information leakage or serious data utility loss. In this paper, we propose a joint multimodal desensitization framework to deal with face-related sensitive information in multimodal data. By utilizing Large Language Models (LLM) and YOLO World, we can accurately identify sensitive facial regions in multimodal data. We propose a customized face-swapping approach based on Stable Diffusion and IP Adapter to achieve visual anonymity, coupled with a Variational Autoencoder (VAE) to process text reconstruction. In addition, CLIP-based constraints are used to ensure the semantic consistency of multimodal data. The experimental results show that the proposed approach can reduce the Re-ID rate to 2.1% with high data utility.

Keywords: Multimodal Desensitization, Cross-Modal Consistency, Privacy Preservation.

1. Introduction

With the rapid growth of data-driven applications, multimodal information such as text, images, and audio [1] is increasingly used in AI training and governance scenarios, including content review, statistical analysis, intelligent surveillance, and smart-city management [2,3]. However, these data sources often contain sensitive personal information, particularly facial cues in images and identity-related descriptions in text [2,4,5]. If disclosed or misused, such information may threaten personal privacy and even public security. Traditional desensitization approaches, such as image blurring or text redaction, still show clear limitations. They typically process each modality separately and therefore fail to capture hidden privacy links across modalities—for instance, when a textual name can still reveal the identity behind an anonymized face [6]. In addition, excessive masking may damage the structure and usability of the original data, reducing its value for downstream AI applications [3].

In order to solve the aforementioned problems, a joint desensitization method based on face swapping and cross-modal information complement is proposed in this paper. The sensitive facial information in the text, such as names, identity expressions, or attributes such as "a man in red clothes," is used to guide the Stable Diffusion model with the IP Adapter to generate the facial areas at the pixel level while keeping the backgrounds and scenes. Meanwhile, a VAE-based text

DOI: 10.54254/2753-8818/2026.32588

reconstruction module is used to update the text information to conform to the modified image [3]. The difference from the traditional operation of masking the sensitive facial information in the image is the use of generative reconstruction to ensure the image reconstruction. Besides, the constraints based on the consistency of the anonymized image and text generated by the CLIP model are introduced to ensure the semantic consistency of the anonymized image and text, which prevents the privacy leakage of the image and text in the cross-modal case while ensuring the usability of the data in the applications of AI model training, text moderation, and image moderation

2. Related work

Data privacy protection has expanded from the old-fashioned safeguarding of structured data to the newer and more complicated sanitizing of various types of data [3,7]. In the structured domain, Fouad et al. proposed the differential privacy-based SDPP algorithm [7], and Yang et al. introduced MIAE (Mutual Information Autoencoder) to strike a balance between privacy protection and data utility [3]. Visual domain: Yu et al. proposed the iPrivacy framework based on multi-task learning for privacy object detection [2], Kuang et al. did research on face anonymization via attention distraction [5]. Text is usually desensitized with named entity recognition and keyword replacement, but these techniques do not always catch implicit interdependencies among different modalities [6]. Generative models have exhibited considerable promise in visual desensitization [4,5,8],yet most current methods still handle text and images independently, which makes it hard to maintain semantic consistency after desensitization.

To address this issue, this study proposes a new framework that can optimize both the face-swapping and text-reconstruction tasks at the same time, and also imposes a CLIP-guided cross-modal consistency constraint. The new framework is based on Stable Diffusion algorithm for face swapping and VAE algorithm for text reconstruction, and also uses YOLO world and LLM prompts to take advantage of both modalities. Compared to the current face swapping algorithms which depend solely on one modality, the newly proposed framework is more likely to guarantee privacy protection and semantic consistency of the text-image pairs in the desensitized data, thus providing a stronger solution for multimodal data desensitization.

3. Proposed methodolgy

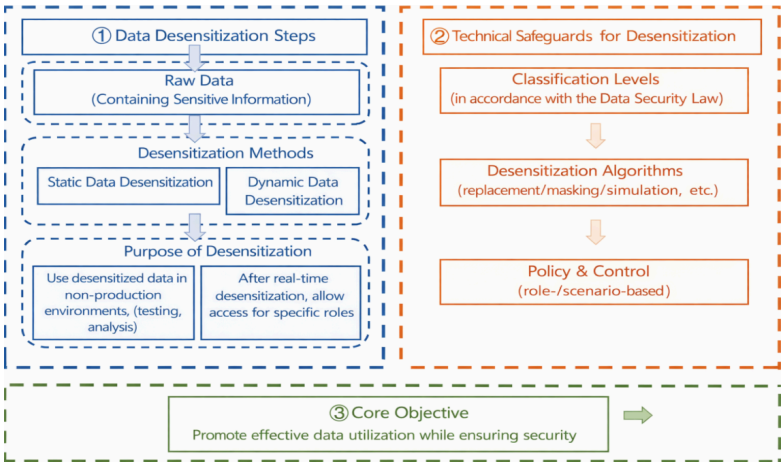

Figure 1. Workflow of vision-language guided face localization utilizing LLM-extracted features and YOLO World

DOI: 10.54254/2753-8818/2026.32588

3.1. Multimodal sensitive information extraction

As illustrated in Figure 1, the procedure starts with extracting face-related attributes from the input text through an LLM and converting them into semantic queries used for vision–language guided face localization. The extracted information—such as gender, age, or other identity-related cues—is then encoded by a pre-trained CLIP text encoder and mapped into a high-dimensional feature representation V_{text} , which captures the semantic characteristics of the sensitive attributes.

3.2. Vision-Language guided localization

We employ the YOLO World architecture, specifically utilizing its Vision-Language Path Aggregation Network (PAN). By treating the semantic vector V_{text} as a query, the model performs cross-modal alignment to predict the spatial region of the sensitive identity. The output is a precise binary face mask MMM, defining the pixel-level coordinates for subsequent processing

3.3. Customized face-swapping with stable diffusion

As illustrated in Figure 2, IP-Adapter and InsightFace are integrated into a Stable Diffusion framework to perform controllable face-swapping for visual anonymization. This module performs latent injection of non-sensitive facial features guided by a target identity. By utilizing the previously generated mask MMM for constrained denoising, the model ensures that the newly synthesized face maintains the original scene's lighting, pose, and background consistency while effectively desensitizing the identity

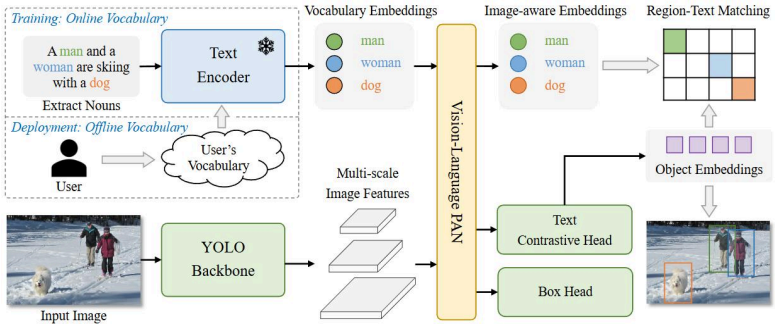

Figure 2. Detailed pipeline of the customized face-swapping algorithm integrated with Stable Diffusion and IP-Adapter

3.4. Vae-guided joint text desensitization

To synchronize the text with the new visual identity, visual features $V_{face}^{'}$ are injected into the latent space of a VAE text encoder. The reconstruction is formulated as:

$$z = E_{text}(T) \oplus \phi(V_{face}^{'})$$

This ensures the text matches the modified image (e.g., changing a specific name to a general description).

DOI: 10.54254/2753-8818/2026.32588

3.5. Cross-modal consistency constraints

To prevent semantic drift, we jointly optimize three loss functions:

$$L_{total} = \lambda_1 L_{img} + \lambda_2 L_{text} + \lambda_3 L_{CLIP}$$

where $LCLIP$ minimizes the cosine distance between the swapped image and desensitized text features inthe CLIP latent space. Additional adaptations include cross-modal knowledge guidance that aligns textextracted face sensitive features with image face features, and enhanced consistency constraints thatiteratively optimize the models if similarity falls below threshold. This joint optimization guaranteesboth privacy and utility preservation [9].As illustrated in Figure 3, the training pipeline coordinatesimage reconstruction loss, text reconstruction loss, and CLIP cross-modal alignment loss in a unifiedframework, allowing visual features from the swapped face to directly guide the VAE text encoder while CLIP enforces semantic coherence between the modified image and the updated text description.

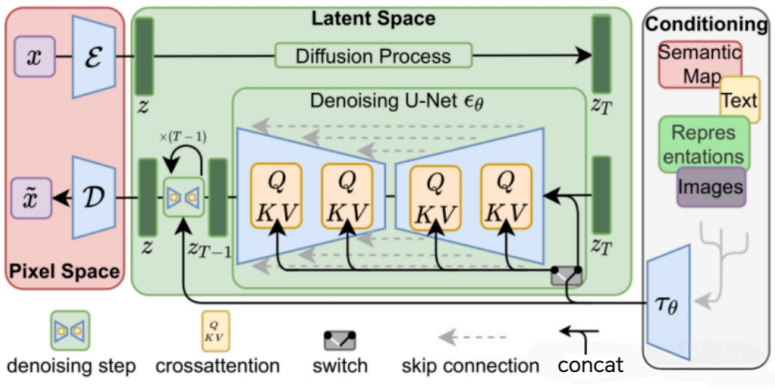

Figure 3. Joint loss optimization and CLIP cross-modal alignment

4. Experiments and analysis

4.1. Experimental setup

We validated the framework using MS COCO and VoxCeleb datasets. Performance was measured by the Re-identification (Re-ID) rate for privacy, and BERTScore (text) and FID (image) for utility.

4.2. Quantitative results

As shown in Table 1, our proposed scheme significantly outperforms traditional methods.

Table 1. Performance comparison of different desensitization methods

Method	Re-ID Rate (↓)	BERTScore (↑)	Image FID (↓)
Original Data	98.2%	1.000	0.00
Gaussian Blur	15.4%	0.762	48.23
Proposed	2.1%	0.954	11.85

DOI: 10.54254/2753-8818/2026.32588

4.3. Qualitative discussion

As shown in Figure 4, the swapped faces blend naturally with the original scene and appear more realistic than those produced by traditional desensitization methods. The textual updates effectively remove specific identifiers while maintaining the core narrative.

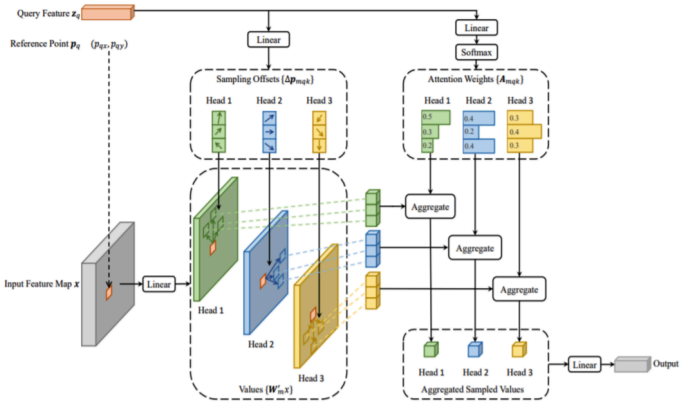

Figure 4. Qualitative comparison of desensitization performance between traditional methods and our proposed joint scheme

The framework was validated on MS COCO (text-image pairs) and VoxCeleb (identity verification) datasets. Ablation studies confirm that removing the CLIP cross-modal module drops implicit privacy detection by 34%, validating the necessity of joint optimization. Visual results show swapped faces are naturally blended with no artifacts, and text updates are grammatically correct and semantically consistent [7]. As shown in Figure 5, the left panel compares original and desensitized image-text pairs, while the right panel shows ablation curves under different λ settings, highlighting the contribution of cross-modal consistency to the privacy-utility trade-off.

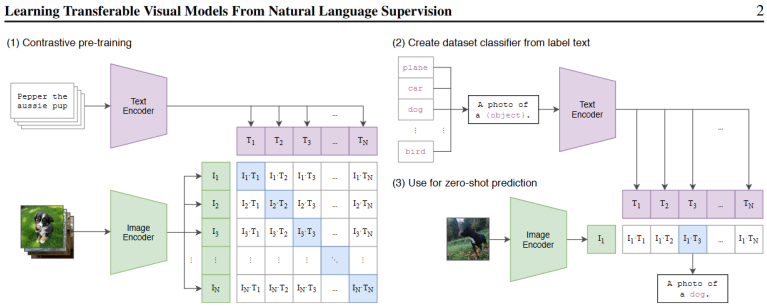

Figure 5. Ablation study and qualitative results

5. Conclusion

This study puts forward a multimodal desensitization method to correct the shortcomings of traditional privacy protection approaches. In the actual world of data, sensitive info is scattered all over different data sources, including visuals. A framework has been put forth for a better way to deal with the issue of leaking sensitive information, which comprises three significant parts: the identification of multimodal sensitive information, generative reconstruction, and cross-modal consistency learning.

DOI: 10.54254/2753-8818/2026.32588

The suggested framework applies the stable diffusion procedure for desensitizing sensitive data in the visual modality. In the textual modality, the proposed framework uses a VAE-based reconstruction module to anonymize sensitive information while keeping its semantic meaning intact. Also, the proposed framework uses CLIP-based feature alignment so that the desensitized sensitive information will be consistent in the textual modality.

The suggested framework prevents the leak of sensitive info but still keeps it useful. The proposed framework gets the right mix of safeguarding personal identities and making the data useful...

Acknowledgments

This work was supported by the research project on cross-modal information completion enabled multimodal data desensitization mechanism. The authors would like to express their sincere gratitude to the School of Mathematical Sciences, Ocean University of China, for providing the necessary research facilities and computing resources. Special thanks are also extended to the project team for their foundational contributions to the VAE, Stable Diffusion, and CLIP consistency constraint framework. We sincerely thank the anonymous reviewers for their helpful comments and suggestions, which have helped improve the quality of this paper.

References

[1] Qian J, Du H, Hou J, et al. Speech sanitizer: Speech content desensitization and voice anonymization. IEEE Transactions on Dependable and Secure Computing, 2021, 18(6): 2631-2642.
[2] Yu J, Zhang B, Kuang Z, et al. iPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. IEEE Transactions on Information Forensics and Security, 2017, 12(5): 1005-1016.
[3] Yang Q, Wang C, Yuan H, et al. Approaching the Information-Theoretic Limit of Privacy Disclosure With Utility Guarantees. IEEE Transactions on Information Forensics and Security, 2024, 19: 3339-3352.
[4] Zhang Y, Ji J, Wen W, et al. Understanding Visual Privacy Protection: A Generalized Framework With an Instance on Facial Privacy. IEEE Transactions on Information Forensics and Security, 2024, 19: 5046-5059.
[5] Kuang Z, Yang X, Shen Y, et al. Facial Identity Anonymization via Intrinsic and Extrinsic Attention Distraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 12406-12415.
[6] Gong T, Wang J, Zhang L. Cross-modal semantic aligning and neighbor-aware completing for robust text-image person retrieval. Information Fusion, 2024, 112: 102544.
[7] Fouad M R, Elbassioni K, Bertino E. A Supermodularity-Based Differential Privacy Preserving Algorithm for Data Anonymization. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7): 1591-1601.
[8] Huang C, Chen S, Zhang Y, et al. A robust approach for privacy data protection: IoT security assurance using generative adversarial imitation learning. IEEE Internet of Things Journal, 2022, 9(18): 17089-17097.
[9] Li N, Li T, Venkatasubramanian S. Closeness: A New Privacy Measure for Data Publishing. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(7): 943-956.