Who Controls the Authorization? Invertible Networks for Copyright Protection in Text-to-Image Synthesis

Supplementary Material

6. Threat Scenario and Authorization Protocol

With the increasing prevalence of generative copyright infringement, protecting image ownership faces more severe challenges and increasingly complex real-world demands. We model a threat scenario involving three parties: (i) a *copyright owner*, who may be the individual depicted in the image or the creator of an artwork, and seeks to publish the image securely online; (ii) a *protection service provider*, who processes the image upon the owner's request to ensure it cannot be misused or repurposed without authorization once published on the internet; (iii) an *attacker*, who downloads the image from online sources and exploits it for unauthorized and potentially malicious purposes.

In this threat scenario, we assume that the protection service provider receives requests from individuals who require privacy-preserving protection. Based on the original image and a user-specific copyright watermark provided by the owner, the service provider performs proactive anti-personalization processing. The resulting copyrighted image, along with a corresponding key, is then returned to the copyright owner. The owner can safely publish or use the copyrighted image, knowing that unauthorized personalization is effectively deterred. The trained invertible neural network (INN) model is securely stored on a cloud server managed by the service provider. This separation of varying recovery data follows a design principle similar to that of zero-watermarking.

As described in the main text, we consider two potential infringement strategies for digital images in networked environments. The first involves direct misuse of the copyrighted image. With our proposed method, the owner can initiate copyright verification, where the recovered watermark serves as strong evidence of unauthorized use. The second strategy involves the unauthorized generation of highly realistic personalized images for more precise and stealthy infringement. Our method effectively prevents this threat by ensuring that any images generated from such unauthorized personalization attempts exhibit visible watermark artifacts, thereby exposing the infringement and significantly limiting their usability.

When a third party seeks authorized personalization for legitimate purposes, an authorization protocol is required to ensure secure and auditable access. The requester must first submit an application to the copyright owner, specifying the intended use and target identity. Upon approval, the owner issues a key and a digital certificate. The certificate enables the requester to contact the protection service provider, who verifies the authenticity of the certificate and grants controlled access to the INN model. Using the authorized key, the requester can reconstruct the high-fidelity restored image via the model's backward process. This image can then be used for authorized and privacy-preserving personalization. All access and authorization records are logged for traceability, ensuring that only certified users can perform personalization.

Dataset	Defense?	FID ↑
ImageNet (general)	×	77.18
	✓	**376.48**
WikiArt (stylization)	×	146.73
	✓	**325.23**

Table 6. Performance of our method on ImageNet dataset for general task and WikiArt dataset for style transfer task.

7. Implementation Details of the Model

Our model comprises a series of inverse compound coupling modules, with 24 modules used in the experiments. This core inverse information exchange module links x_{ori} and x_{w} through four transformations: $\rho(\cdot)$, $\psi(\cdot)$, $\varphi(\cdot)$, $\eta(\cdot)$. Among them, $\rho(\cdot)$, $\varphi(\cdot)$,and $\eta(\cdot)$ are implemented using 5-layer residual dense blocks, while $\psi(\cdot)$ incorporates SE-inspired pooling-linear layers to enhance frequency interaction.

8. Generalization Performance

To evaluate the generalizability of our method, we perform personalization tasks on the general object dataset ImageNet and style transfer tasks on WikiArt, using 20 object categories and 10 painting styles, respectively.

As shown in Table 6, the FID scores of the images generated by diffusion increase significantly under both tasks after applying our defense, indicating that the generation quality is effectively degraded, thus preventing unauthorized personalization. The subjective results are visualized in Figures 8 and 9, respectively.

9. Evaluation on Textual Inversion

We further evaluate our method on textual inversion infringement to demonstrate its robustness against diverse personalization attacks. Textual inversion is another impor-

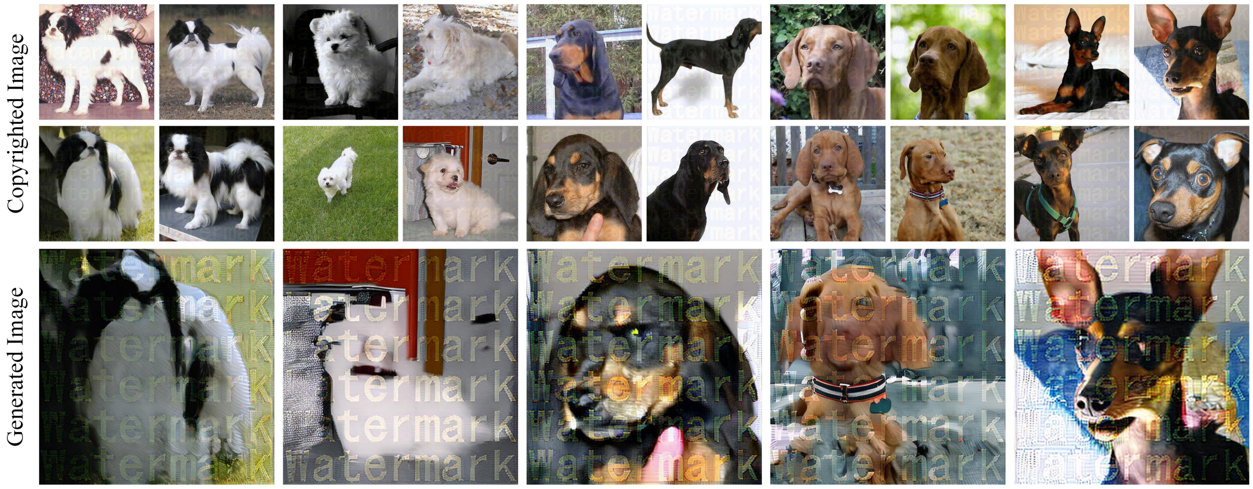

Figure 8. Visual results on ImageNet personalization task.

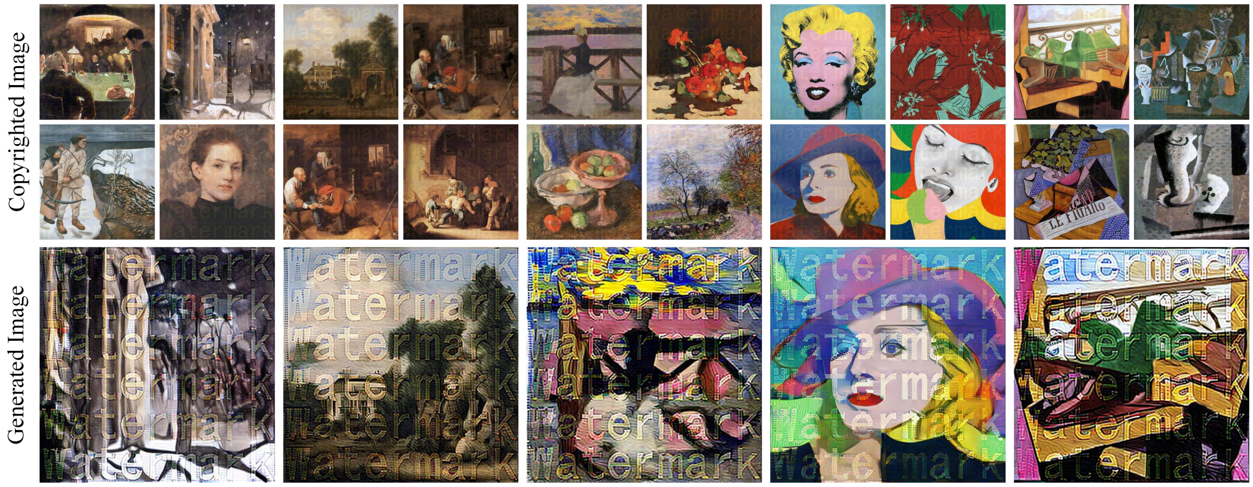

Figure 9. Visual results on WikiArt stylization task.

tant method that learns to represent the concept of images in a pseudo-token, which is then used to guide image generation. As shown in Table 7, our approach consistently maintains defense performance, enabled by a VAE encoder-based protection mechanism and a contrastive learning-driven optimization strategy.

10. More Visualization Results

10.1. Ablation Study

As shown in Figure 10, removing our modules produces generated images that still retain watermark traces, but with reduced visibility and clearer facial feature details, indicating weakened protection. Notably, removing the DWT module results in smoother outputs that resemble the degradation pattern of the copyrighted image. These modules are essential for effective copyright protection.

Method		FID ↑	FDFR ↑	ISM ↓
No Defense		145.82	0.0743	0.5123
Defense	ASPL	243.56	0.3126	**0.3312**
	FSMG	218.82	0.2745	0.3561
	MetaClock	199.54	0.1532	0.4164
	Wm-AdvDM	183.57	0.0906	0.4478
	Ours	**246.08**	**0.3175**	0.3871

Table 7. Comparing the defense performance of current state-of-the-art methods.

10.2. Inference with Different Text Prompts

Our method maintains robust anti-personalization capability even when tested with unseen text prompts during inference. As shown in Figure 12, this indicates that our

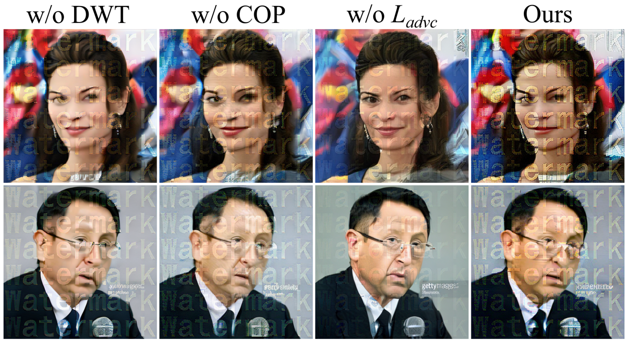

Figure 10. Qualitative results of ablation study.

copyright protection framework establishes a strong association between the copyright watermark and the target token "sks". Additionally, compared to complex facial features, the simpler structure of the watermark is captured more effectively by the diffusion model. As shown in Figure 13, other methods exhibit varying degrees of performance degradation when the prompt is altered, our method maintains consistent protective performance consistently across varying text prompts.

10.3. Results of Visual Quality

Figure 14 shows subjective results of copyrighted images generated by comparison methods, providing visualizations that demonstrate the impact of adversarial perturbations on image quality. We advised to zoom in for a detailed examination. While these perturbations do not significantly alter the overall semantics or prominent facial features of the images, they introduce noticeable alterations to fine-grained textures, potentially limiting the images' applicability in high-resolution tasks. In contrast, the copyrighted images produced by our method, once authorized, maintain remarkable similarity to the original images, with no perceptible artifacts or distortions.

10.4. Results of Robustness

To demonstrate the robustness of our method against adversarial purification techniques, Figure11 presents copyrighted images before and after purification, along with their corresponding generated images. As shown, Gaussian blurring smooths image edges, significantly weakening copyright protection, as evidenced by the clearer facial features in the generated images. However, JPEG compression and super-resolution (SR) have negligible impact, preserving discernible copyright watermark traces in the generated samples.

11. Settings of Other Models

We use set A from Anti-DreamBooth to train all baseline methods. Reconstructing Wm-AdvDM based on the descriptions provided in its paper, we found that the original settings caused the network to fall into a local optimum early in training, preventing adversarial loss optimization. Therefore, we adjusted the weight ratio of Wm-AdvDM from 1:10 to 8:1, allowing the algorithm to better realize its intended capability.

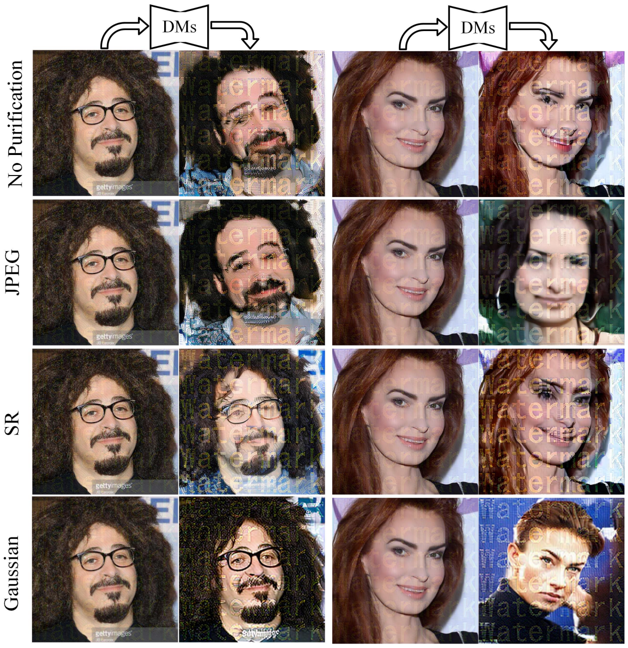

Figure 11. Purified copyrighted images and their generated results

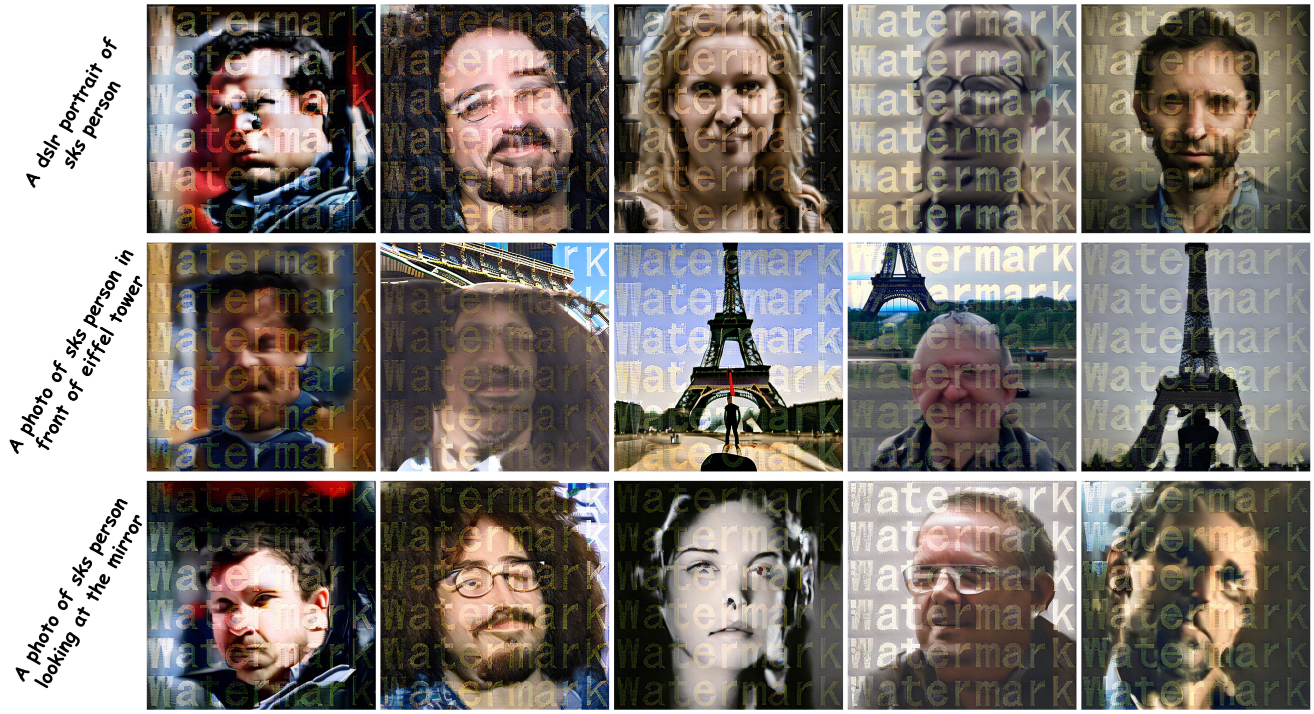
A dslr portrait of sks person
A photo of sks person in front of eiffel tower
A photo of sks person looking at the mirror

Figure 12. Generated results of diffusion models using our copyrighted images under varying text prompts.

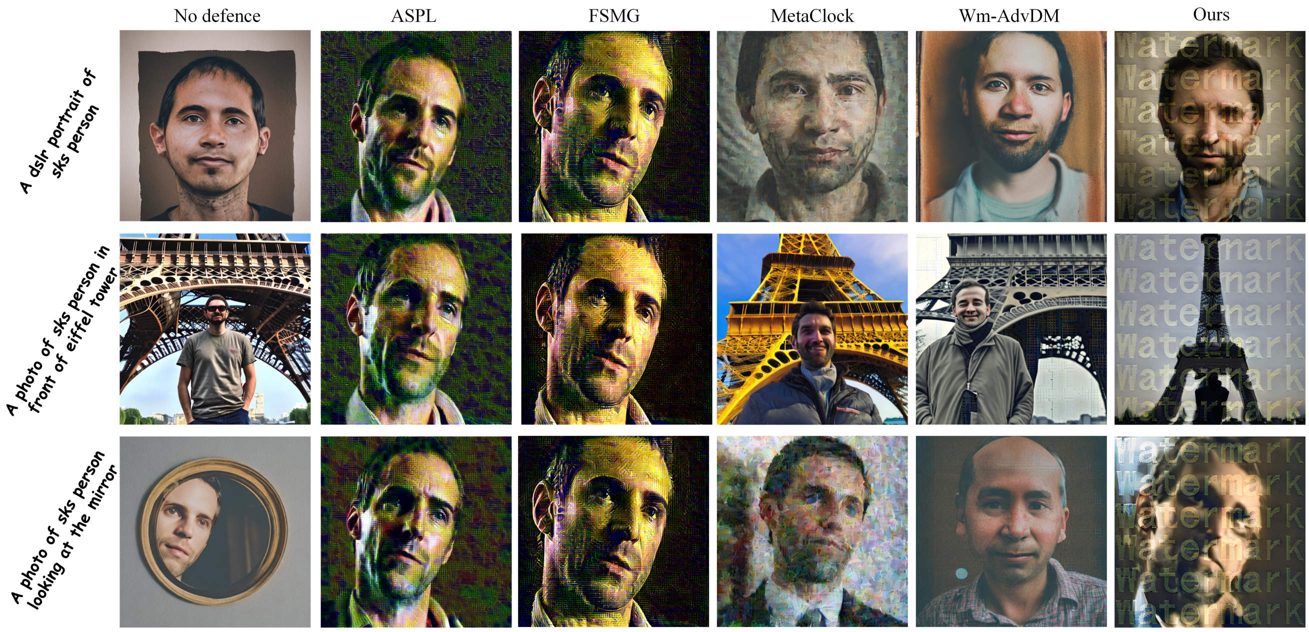
No defence
ASPL
FSMG
MetaClock
Wm-AdvDM
Ours
A dslr portrait of sks person
A photo of sks person in front of eiffel tower
A photo of sks person looking at the mirror

Figure 13. Comparison of personalization generation results across different methods under varying text prompts.

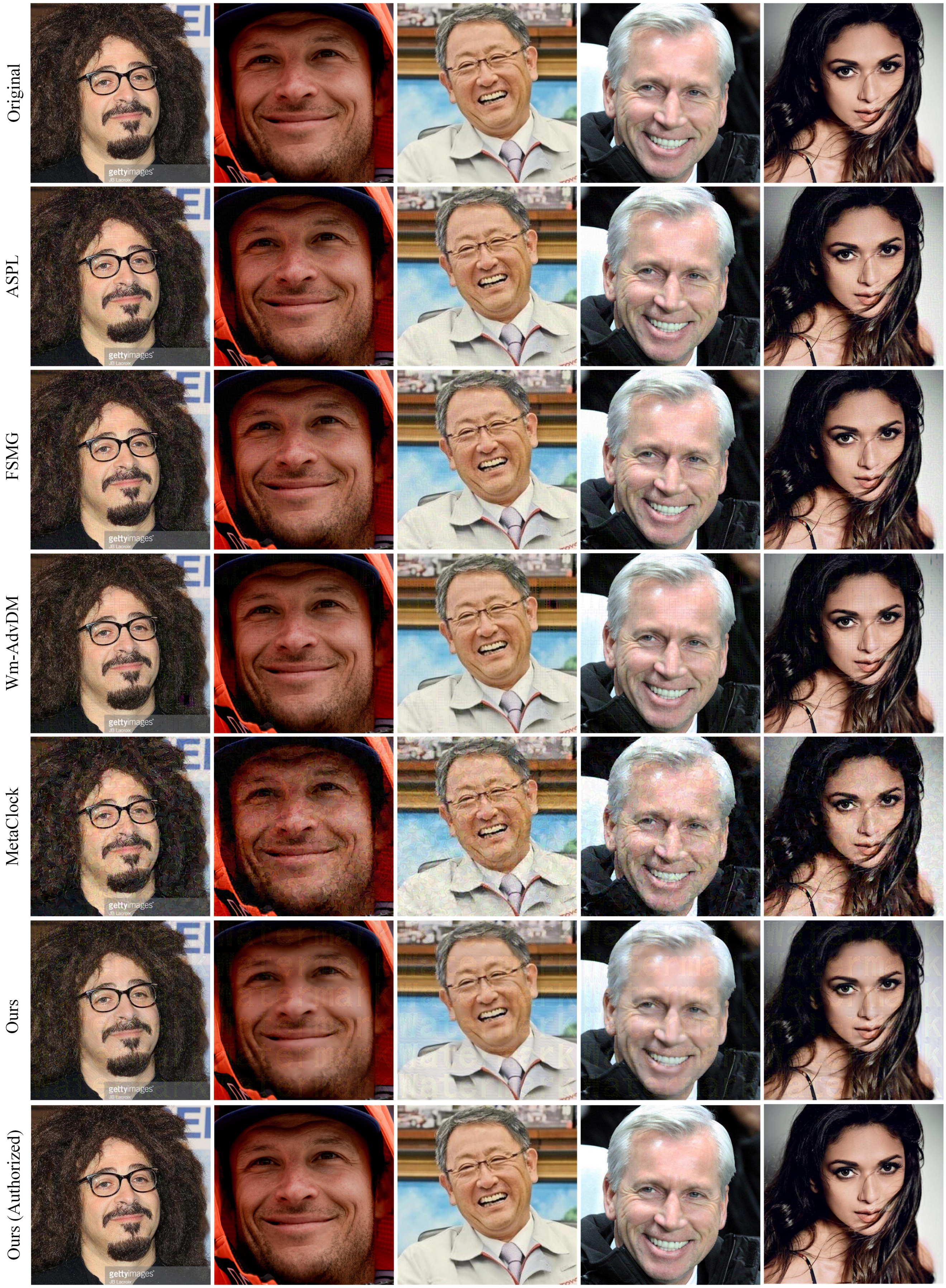

Figure 14. Visual Quality of Copyrighted Images Generated by Different Methods.