1 Expanded Stego Visual Examples

The following are more visual examples with other pictures and other models, such as baseline, all at 256×256 resolution. We observe how visible artefacts due to larger capacity are reduced by metameric-objective training. The same payload sentence is used, truncated by the capability of each setting.

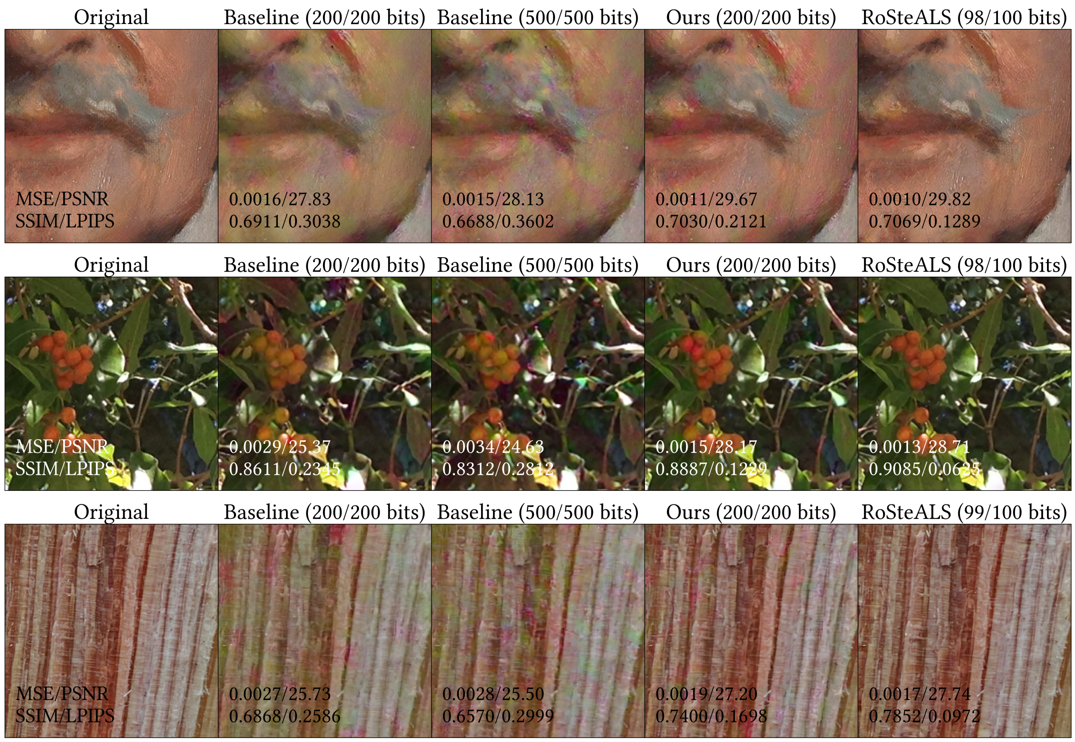

Figure 1: More examples of stego images with their image metrics and payload accuracy in correct/capacity format (Source: MetFaces [Karras et al. 2020] and CLIC [Toderici et al. 2020]).

Table 1: Results of autoencoders evaluation. Remarkable high performance are highlighted in green and low in red.

Series	Version	Downsample Factor	Channels	Encoding Time (ms)	Decoding Time (ms)	MSE	PSNR	SSIM	LPIPS
TAESD [Boer Bohan 2023]	SD	8	4	0.0025	0.0025	0.0019	31.0179	0.7737	0.2162
	SDXL	8	4	0.0020	0.0023	0.0017	32.6519	0.7898	0.2136
	SD3	8	16	0.0019	0.0021	0.0008	35.3964	0.8895	0.1275
OptVQ [Zhang et al. 2024]	16x16x4	16	256	0.0065	0.0069	0.0021	29.5426	0.7963	0.1629
	16x16x8	16	256	0.0064	0.0069	0.0033	28.2396	0.8450	0.1414
SBER-MoVQ [Maltseva et al. 2023]	67M	8	4	0.0048	0.0097	0.0926	11.8707	0.2425	0.5016
	270M	8	4	0.0046	0.0096	0.0847	12.2756	0.2453	0.5076
Taming [Esser et al. 2021]	Z16384	16	256	0.0051	0.0071	0.0053	26.3364	0.6189	0.2943
LDM [Rombach et al. 2022]	VQ F16	16	8	0.0051	0.0069	0.0041	27.5614	0.6876	0.2637
	VQ F8	8	4	0.0044	0.0060	0.0030	29.0924	0.7306	0.1922
	VQ F4	4	3	0.0034	0.0034	0.0012	33.3598	0.8608	0.0906
	VQ F4 noat	4	3	0.0033	0.0033	0.0007	35.2590	0.8988	0.0680
	KL F4	4	3	0.0218	0.0034	0.0007	35.9032	0.8982	0.0818
	KL F32	32	64	0.0244	0.0076	0.0032	29.1994	0.7408	0.2190

2 Autoencoders Assessment for Model's Backbone

We conducted an assessment of state-of-the-art, popular, and open-sourced pretrained autoencoders to select the best backbone candidates. The study spans from autoencoders family (vanilla AE, VAE, VQGAN, etc.) to their specific versions. These models are all independently evaluated under the same framework. In concrete, a mixture of validation set from MetFaces [Karras et al. 2020] and CLIC [Toderici et al. 2020] datasets is used to assess the reconstruction quality of the autoencoders, along with their compression capabilities and speed.

The results of assessment are shown in Table 1. We notice that LDM-VQ-F4, KL-F4, and TAESD-SD3 stand out in image quality, although KL-F4 has significant slower encoding time. So, among the two left, LDM-VQ-F4 (including noat version) are prioritised over TAESD-SD3 due to lower LPIPS But it is worth mentioning that TAESD-SD3 achieves similar level of quality with a much lighter model, a few MB. Therefore, the latter is used for quicker preliminary experiments, while LDM-VQ-F4 is used for the larger ones due to greater potential.

Table 2: Tables of steganographic experiments. Pilot experiments (left) are trained with maximum 100 epochs, patience 10, and learning rate of 8e-5. Hyper-parameter optimization (HPO) experiments (middle) increases patience to 30 and learning rate to 1e-4. Full experiments (right) have uncapped epochs (usually 400-500), where resolution, payload capacity, and image loss function, are shown respectively. First row of each table are reference baselines where following changes are applied respect to.

Pilot Experiments	Bits Acc.	LPIPS
Pilot Baseline	0.5311	0.2554
Augmentation: On	0.6426	0.2302
Train Size: 2K	0.7338	0.3602
Backbone: TAESD3	0.8045	0.4040
MSE weight: 0	0.5861	0.9380
Color Space: YUV	0.5081	0.1218
Resolution: 256	0.5016	0.1211
Hider: RoSteALS	0.5013	0.0456

HPO Experiments	Bits Acc.	LPIPS
HPO Result	0.9999	0.2675
No Augmentation + Mix Sum	0.9999	0.2826
TAESD3 Backbone + Mix Sum	0.9995	0.3868
Batch Size 32	1	0.3026
Mix Sum	1	0.3019

Full Experiments	Bits Acc.	LPIPS
100 128 MSE	0.9999	0.1924
100 128 Metameric	0.9998	0.1409
100 256 MSE	1	0.1621
100 256 Metameric	0.9998	0.1198
200 128 MSE	0.9995	0.2251
200 128 Metameric	0.9991	0.1470
200 256 MSE	1	0.2047
200 256 Metameric	0.9998	0.1288

3 Hyper-Parameter Optimization, Architecture Search, and Ablation Studies

After ensuring that image could be reconstructed appropriately, we conducted many preliminary studies, alternating settings, but they all failed to learn payload embedding. It was only until curriculum learning when complete pipeline could be trained. The most relevant results of these experiments are shown in Table 2. In the pilot studies, we explored many settings from a pilot baseline that uses 1K train images, LDM-VQ-F4 as backbone, MSE loss weights 0.1, minimum resolution of 128×128, and payload capacity of 100. From this study, we find that variety of examples speed up the learning process, whereas more complex patterns may lead to better results but are slow to converge.

After getting a better sense of search space, we conducted a hyper-parameter optimization (HPO) study, where the best settings were found. It is characterized by the method described in the main paper like Conv Sum ("sandwiching conv layers") merger architecture and batch size 8, adding on the positive conditions in pilot study. Finally, full experiments on resolutions and capacity of choice were conducted, showing an expected trend of increase in quality performance when larger resolution or reduced payload capacity. Noticeably, the metameric objective always outperforms the MSE objective, with similar payload accuracy, showing the effectiveness of foveated training in steganography.

References

Ollin Boer Bohan. 2023. Tiny AutoEncoder for Stable Diffusion (TAESD). https://github.com/madebyollin/taesd. commit dc25abb, accessed 2025-05-12.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems* 33 (2020), 12104–12114.

Anastasia Maltseva, Arseniy Shakhmatov, Andrey Kuznetsov, and Denis Dimitrov. 2023. SBER-MoVQGAN. https://github.com/ai-forever/MoVQGAN. commit f3976d8, accessed 2025-06-08.

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

George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Balle, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. 2024, 2022, 2021, 2020. Workshop and challenge on learned image compression (CLIC2020 to CLIC2024). In *CVPR*.

Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. 2024. Preventing local pitfalls in vector quantization via optimal transport. *arXiv preprint arXiv:2412.15195* (2024).