Seeing What You Say: Expressive Image Generation from Speech

Jiyoung Lee[1†*] Song Park[2†] Sanghyuk Chun[3†] Soo-Whan Chung[4]

Ewha Womans University[1] NAVER AI Lab[2] Princeton University[3] NAVER Cloud[4]

Abstract

This paper proposes a unified end-to-end speech-to-image model, ***VoxStudio****, the first attempt to generate an expressive image directly from a spoken description by aligning both linguistic and paralinguistic information. We eliminate the need for an additional speech-to-text module, which often ignores the hidden details beyond text,* e.g.*, tone or emotion. To further advance this direction, we introduce a new emotional speech and image dataset, VoxEmoset, that pairs emotional speech synthesized with corresponding visual scenes, enabling training and evaluation of both semantic and affective aspects of speech-to-image generation. Comprehensive experiments on the SpokenCOCO, Flickr8kAudio, and VoxEmoset benchmarks demonstrate the feasibility of our method and highlight key challenges, including emotional consistency and linguistic ambiguity, paving the way for future research. The project page is* http://mmai.ewha.ac.kr/voxstudio/

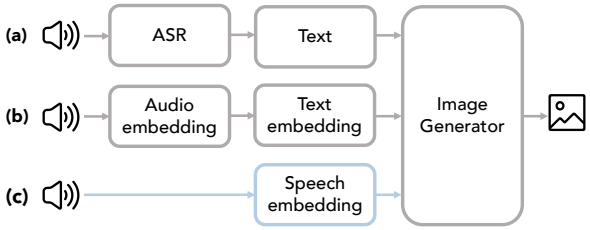
1. Introduction

Imagination (production of sensations or feelings to create mental images) by listening to an explanation (speech) is a natural cognitive process. Speech is a primary and intuitive modality for human communication, capable of conveying both explicit semantics and rich paralinguistic cues such as emotion, tone, and speaker intent. Unlike written language, which abstracts meaning into discrete symbols, speech embodies expressive characteristics that often carry critical information beyond lexical content. These expressive signals offer a promising foundation for cross-modal generation tasks, particularly in generating visual content that aligns not only with the literal meaning of spoken descriptions but also with their affective undertones.

Recent advances in text-to-image (T2I) generation, especially diffusion-based models such as Stable Diffusion [26, 30], have achieved impressive results in terms of image fidelity and semantic alignment. However, these models are intrinsically limited by their reliance on text inputs, which cannot faithfully capture the nuances of emotional intent or vocal expressiveness embedded in speech. To address this, some prior works [43, 44] have introduced emotion-aware image generation by incorporating explicit textual modifiers or sentiment labels. Nevertheless, such methods are still constrained by the semantic ceiling of text, and often fail to account for subtle variations in delivery such as pitch, rhythm, or prosody that are intrinsic to human speech.

Other efforts have explored audio-to-image generation [16, 18, 35, 36] by mapping audio features into a textual embedding space or aligning them with pretrained T2I models. However, they typically disregard the affective content

Figure 1. Samples produced by **VoxStudio** from spoken descriptions.



	(a) Cascaded	(b) Mapping	(c) **VoxStudio**
Time	654.3ms	23.8ms	22.2ms
GFLOPS	4919G	154G	128G
Params.	2.36B	1.2B	0.64B

Figure 2. (a) The cascaded system consisting of ASR ([28]) and T2I ([30]), and (b) audio-to-text feature mapping-based methods [36, 38] limits in cost than (c) **VoxStudio** (ours). The diffusion model is excluded to compute GFLOPs, time computations, Params.

† Partly works done at NAVER AI Lab

* Corresponding author: lee.jiyoung@ewha.ac.kr

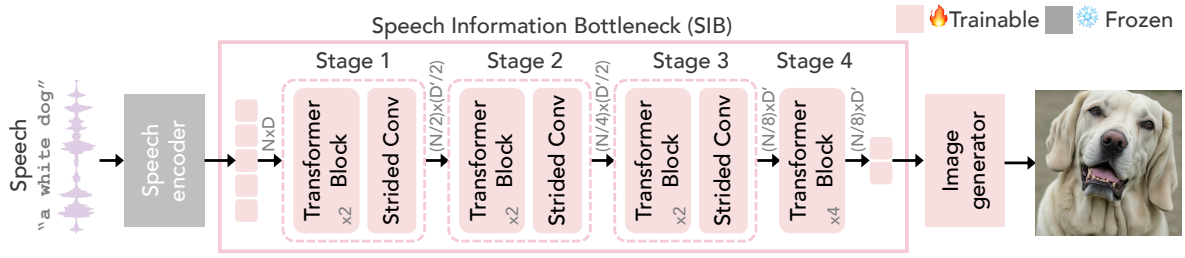


Figure 3. The overall framework of **VoxStudio** for expressive image synthesis from spoken description.

of the audio or reduce it to textual proxies. Moreover, many rely on cascaded frameworks where speech is first transcribed into text via automatic speech recognition (ASR), which introduces latency, transcription errors, and the loss of paralinguistic information. These shortcomings are especially pronounced in under-resourced or unwritten languages – there are over 7,100 languages worldwide [5], yet widely-used ASR services such as Google's Speech Recognition API support a limited number (125) of languages[1] –, where ASR models may not be available, further limiting accessibility and inclusivity.

To overcome these challenges, we propose a novel framework, **VoxStudio**, that directly generates expressive images from speech in an end-to-end manner. Rather than treating speech as an intermediary step toward text, **VoxStudio** embraces it as a fully expressive and sufficient modality for visual content creation. Our model learns to extract and compress meaningful phonetic features, including both semantic and emotional information, with a tailored speech information bottleneck (SIB) module, and conditions a latent diffusion model for image generation. In addition, we introduce VoxEmoset, a large-scale benchmark dataset that pairs emotional speech with corresponding images, allowing for rigorous evaluation of emotional fidelity in generated outputs. By embracing speech as a standalone modality, **VoxStudio** bridges a significant gap in multimodal generative research, laying the groundwork for affective AI systems capable of responding to human expression.

2. VoxStudio

Unlike conventional cascade approaches (Fig. 2 (a, b)), our method is designed to directly integrate speech representations into the image generation process (Fig. 2 (c)). Our approach reduces computational overhead by directly encoding speech features into image space while preserving emotional and semantic fidelity. As shown in Fig. 2, our approach uses less GFLOPS and parameters than conventional approaches, using the proposed SIB to assist in capturing diverse semantic features in image synthesis. Fig. 3 shows the overall framework of **VoxStudio**, consisting of speech encoder, SIB, and image generator.

We consider two pre-trained speech models, SONAR [4] and Whisper large-v3 [28], to deploy comprehensive speech features considering both linguistic and paralinguistic information. It is known to be capable of capturing paralinguistic information, such as emotion and speaker identity [7, 48]. Formally, given an input speech X, we obtain speech embedding $s \in \mathrm{R}^{D \times N}$, where N depends on the length of the speech and models, and D is the channel dimension of the final output layer.

Although speech embeddings contain rich representations, they are excessively long and lead to a lower information density in each speech token compared to text (*e.g.*, Whisper encodes a maximum of 1500 tokens for 30-seconds long speech, while CLIP text encoder makes 77 tokens). This low density and redundancy make direct usage challenging to condition the image generator. To solve this problem, we consider a Transformer-based speech information bottleneck (SIB) module. SIB compacts semantics in speech embeddings, motivated by previous works [11, 37] applied to image and audio encoders. As shown in Fig. 3, SIB reduces the number of speech tokens with a strided convolution layer after a Transformer block along the time axis. Based on our findings, a pooling ratio of 8 provides the optimal balance, allowing us to maximize the information retention of speech features. As a result, the initial embedding s is processed into a compressed speech condition $c = f_\psi(s)$, where $c \in \mathrm{R}^{M \times D'}$, $M = N/8$ and D' is the input channel of the cross-attention block in the image generator. By leveraging such compressed representations, our method improves the efficiency of speech-to-image while preserving both linguistic and emotional expressiveness.

The image generator $\epsilon_\theta(\cdot)$ is based on latent diffusion model [30] which has demonstrated remarkable fidelity compared to GAN-based methods widely used in early audio-to-image works [40, 41]. The speech condition c, compressed through SIB, is fed into $\epsilon_\theta(\cdot)$ to guide the synthesis process. Specifically, the speech embeddings are injected into the UNet through cross-attention layers to condition the image synthesis. This conditioning allows the model to incorporate the emotional, semantic content of speech into the generation process. The image generator and SIB are optimized with the diffusion loss [30]. Finally, the denoised latent is decoded into the image through the decoder [19]. We note that our framework has no specific

[1] https://cloud.google.com/speech-to-text

Figure 4. Examples from VoxEmoset (expressive tone).

Benchmark	# Images	# Speech	Emotion	ClipScore	NMOS
SpokenCOCO [14]	123k	615k	✗	30.42	2.9616
Flickr8kAudio [9]	8k	40k	✗	31.27	2.9689
VoxEmoset (ours)	82k	247k	✓	30.27	2.9683

Table 1. Comparison of existing datasets and VoxEmoset.

design choices for the image generator.

Since training image generator anew requires a vast amount of resources, we initialize the image generator with a pre-trained T2I model for efficient learning. Only the diffusion loss $\mathcal{L}_{\text{diff}}$ is used for optimizing the parameters θ of the generator and ψ of the SIB, and we do not design a specialized loss function (*e.g.*, contrastive learning in [38], AR moedling in [18]) for speech-to-image learning. Our simple training framework ensures versatile connections for various image generators. More implementation details are in Appendix B.1.

3. VoxEmoset Benchmark

One major challenge of **VoxStudio** is the lack of well-designed large-scale image-speech paired datasets, which contain delicate information beyond text, *e.g.*, tone or emotion. To demonstrate the benefit of directly learning from speech, we build a new large-scale image-speech paired benchmark **VoxEmoset**. Previous benchmarks [10, 14] often overlooked paralinguistic features in speech. Moreover, prior datasets required significant costs for human recordings, limiting their scalability.

Our benchmark uses the partial of 118k image subset in EmoSet [42], emotional image dataset annotated with Mikels model [25]. In line with [6, 23, 33], we group amusement and excitement into a unified category, 'enjoyment', and exclude 'awe' and 'contentment' categories because they are hard to distinguish in voice. The final number of images in VoxEmoset is shown in Tab. 1 and Tab. B1.

We then generate captions using the instruction prompt in Appendix A, restricting emotional expressions and focusing on factual descriptions instead. We use LLaVA-OneVision [20] to generate three different captions are produced for each image to prevent the model from simply generating emotionally biased captions. These captions are subsequently converted into speech using F5-TTS [3] guided by emotion reference signals from speech emotion recognition (SER) datasets including CREMA-D [1], MEAD [39], and RAVDESS [21].

We validate emotional intensity using Emotion2Vec [24], filtering and re-generating inadequate samples. To objectively assess speech quality, we randomly sample 10k speech clips from each dataset and measure NMOS [29]. Tab. 1 shows that VoxEmoset is compatible with existing datasets in terms of speech quality (NMOS) and description quality (CLIPScore). However, only our benchmark explicitly expresses emotion in speech, as shown in Fig. 4. Further details are provided in Appendix.

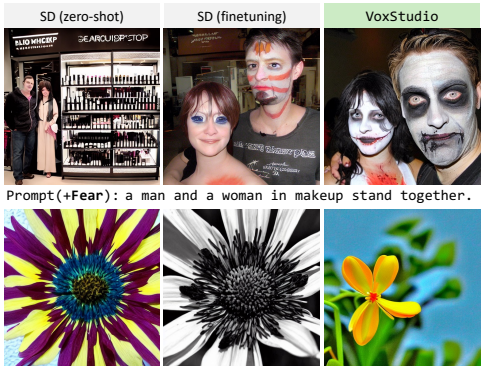


Figure 5. Qualitative comparison between SD using text prompts and **VoxStudio** using speech prompts.

4. Experiments

4.1. Experimental setup

Datasets. We use SpokenCOCO [14] and VoxEmoset to train **VoxStudio**. Flickr8kAudio [9] is used to evaluate zero-shot generalizability. Each image in SpokenCOCO and Flickr8kAudio has five voice recordings from unskilled annotators, resulting in inherently noisy audio (*e.g.*, the recording may contain background noise, reading speed or volume can vary, and pronunciation may not be as clear as that of skilled voice actors as in Appendix B.5). VoxEmoset is automatically generated and less prone to recording noise. We use the Karpathy split [17] for SpokenCOCO.

Evaluation metrics. Following previous works [32, 47], we evaluate the generation quality using FID [13], while content alignment between speech and generated images is measured with CLIPScore [12] using text transcriptions of speech. For SpokenCOCO and VoxEmoset, random samples of 10k condition prompts, either speech or text, are used for evaluation. For Flickr8kAudio, we use 5k test prompts for evaluation. We also report emotion classification accuracy (Emo-A) [43] on generated images to examine whether the results reflect emotion from prompts. Note that we measure accuracy only with scores for the 5 emotion categories – *'amusement'* and *'excitement'* are classified as the same class– in the trained classifier [43].

Method	Training Data	Input	(Spoken)COCO		VoxEmoset		
			FID↓	CLIPScore↑	FID↓	CLIPScore↑	Emo-A↑
SD1.5 [30]	-	T	23.37	31.14	20.21	31.70	60.81
Whisper [28] (ASR) (+SD1.5)	-	T	22.95	31.08	20.23	31.57	60.41
SD1.5 (Finetuning)	SpokenCOCO, VoxEmoset	T	**22.45**	**31.77**	**18.31**	**31.72**	**69.38**
SpeechCLIP+ [38] (+SD1.5)	SpokenCOCO, Flickr8kAudio	S	28.29	25.03	33.75	21.84	37.42
TMT [18] (+SD2.1)	SpokenCOCO, Flickr8kAudio †	S	25.48	28.26	29.48	26.08	48.54
VoxStudio	SpokenCOCO	S	**24.95**	**29.04**	32.60	26.16	46.20
VoxStudio	SpokenCOCO, VoxEmoset	S	27.15	27.27	**19.94**	**29.04**	**71.70**

Table 2. Performance comparison with baselines [18, 30, 38]. 'Input' denotes the data type of input condition for generative models: 'T' is text and 'S' is speech. †: TMT [18] used an additional 15M synthesized speech for training.

Method	Training	FID↓	CLIPScore↑
SpeechCLIP+ [38]	✓	63.19	23.71
TMT [18]	✓	57.34	26.98
VoxStudio	✗	**55.80**	**29.60**

Table 3. Performance comparison on Flikr8kAudio [9]

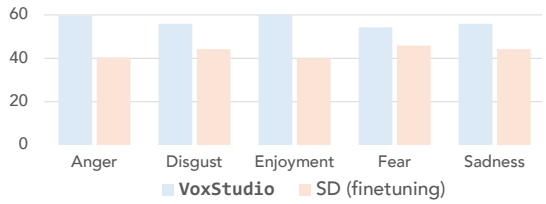


Figure 6. Human evaluation of emotion consistency.

4.2. Results

Results on SpokenCOCO and VoxEmoset. Tab. 2 shows the comparison of **VoxStudio** and baselines on SpokenCOCO and VoxEmoset. SD1.5 with the text inputs (*i.e.*, without speech) is shown as a baseline. Especially, Fig. 5 highlights the stark contrast between text- and speech-based generation. While speech conveys emotions even with the same wording, the text-based model inherently ignores these cues and focuses on fact-based generation. Even when trained on VoxEmoset, 'SD (finetuning)' struggles to express emotions as semantic content, but speech leads to a more rich and intense emotional expression. Moreover, as shown in the last example in Fig. 5, the CLIP encoder [27] often overlooks information from the latter part of a sentence [46] (*e.g.*, 'bright yellow' in the last example). However, **VoxStudio** excels in conveying emotions when trained on the same datasets. This advocates that speech, as a richer modality for emotional expression, provides a more effective signal to generate emotionally compelling images.

Remarkably, **VoxStudio** outperforms SpeechCLIP+ and TMT on SpokenCOCO, where **VoxStudio** does not use Flickr8kAudio for training. While TMT additionally used huge synthesized speech data from CC3M [34] and CC12M [2] for training, **VoxStudio** also show comparable results on VoxEmoset. This result demonstrates that our diffusion model is a powerful learner for speech-to-expressive image alignment than contrastive learning [38] and auto-regressive training [18]. The qualitative comparison on SpokenCOCO also shows that SpeechCLIP+ and TMT often ignore keywords in the prompts, while **VoxStudio** can capture the details, as shown in Fig. B2a.

Results on Flickr8kAudio. Tab. 3 shows the performance comparison on Flickr8kAudio. Here, while TMT and SpeechCLIP+ used Flickr8kAudio for training, **VoxStudio** was evaluated in a zero-shot manner. Surprisingly, **VoxStudio** outperforms existing methods by large margins. It illustrates that end-to-end training in **VoxStudio** is more robust in aligning the speech-language space. By contrast, speech features in **VoxStudio** are more robust to the order or length of the prompt. Moreover, VoxEmoset might improve the robustness on generality as shown in Fig. B2b.

Human evaluation. We also conducted a user study. 26 participants evaluated 25 images to rate how well the emotion conveyed in the image matched the given speech. Specifically, **VoxStudio** is compared with a text-prompted SD1.5 model finetuned on the same datasets, asking which generated images better aligned with the emotion in speech. Fig. 6 shows the results from **VoxStudio** are more aligned with the emotion than SD in all categories, highlighting the effectiveness of speech prompts for expressive image synthesis.

Analysis. To analyze **VoxStudio**'s robustness of design choices and VoxEmoset's impact, results of same description with different emotions, fine-tuning method and scale for generator, effectiveness of speech embedding, and utilization in image editing are provided in Appendix. More qualitative results are also illustrated in Fig. B7 and Fig. B8.

5. Conclusion

VoxStudio is the first end-to-end speech-to-image generation model that totally leverages speech's expressiveness to generate emotionally aligned images. VoxEmoset is a new benchmark, built cheaply, but complementary with real-world datasets. Experiments demonstrated that **VoxStudio** not only outperforms prior speech-based methods in conveying sentiment through images, but also matches text-driven approaches in semantic alignment.

References

[1] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 3, 7

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 4

[3] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024. 3, 7

[4] Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv e-prints*, pages arXiv–2308, 2023. 2, 7, 8

[5] David M. Eberhard, Gary F. Simons, and Charles D. Fennig. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, twenty-sixth edition, 2023. 2

[6] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. 3

[7] Erik Goron, Lena Asai, Elias Rut, and Martin Dinov. Improving domain generalization in speech emotion recognition with whisper. In *ICASSP*, 2024. 2

[8] Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts. *arXiv preprint arXiv:2406.09162*, 2024. 10

[9] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on ASRU*, 2015. 3, 4, 7

[10] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *NeurIPS*, 2016. 3

[11] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *CVPR*, 2021. 2

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 3

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 3

[14] Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. In *ACL*, 2021. 3, 7

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 8

[16] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *ICCV*, 2023. 1

[17] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*, pages 1889–1897, 2014. 3

[18] Minsu Kim, Jee-weon Jung, Hyeongseop Rha, Soumi Maiti, Siddhant Arora, Xuankai Chang, Shinji Watanabe, and Yong Man Ro. Tmt: Tri-modal translation between speech, image, and text by processing different modalities as different languages. *arXiv preprint arXiv:2402.16021*, 2024. 1, 3, 4, 8

[19] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 7

[21] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 3, 7

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7

[23] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Interspeech*, 2024. 3, 7

[24] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *ACL Findings*, 2024. 3

[25] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37:626–630, 2005. 3

[26] 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. *arXiv preprint arXiv:2307.01952*, 2023. 1, 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PmLR, 2021. 4

[28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023. 1, 2, 4, 7, 8

[29] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP*, 2022. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 4, 8