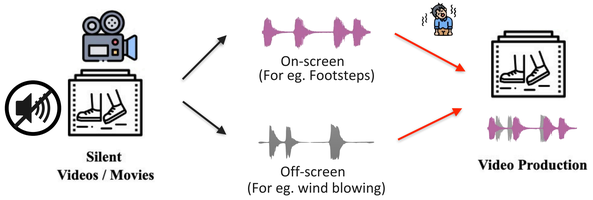
VinTAGe: Joint Video and Text Conditioning for Holistic Audio Generation

Saksham Singh Kushwaha, Yapeng Tian
The University of Texas at Dallas

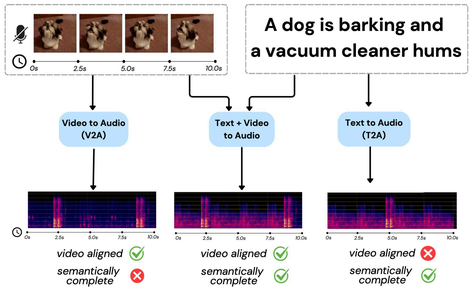
Project Page

Video-to-audio generation requires manually effort



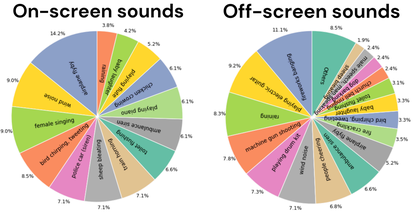
Complex post-production audio requires specialized mixing skills.

Holistic audio generation: Text + Video → Audio

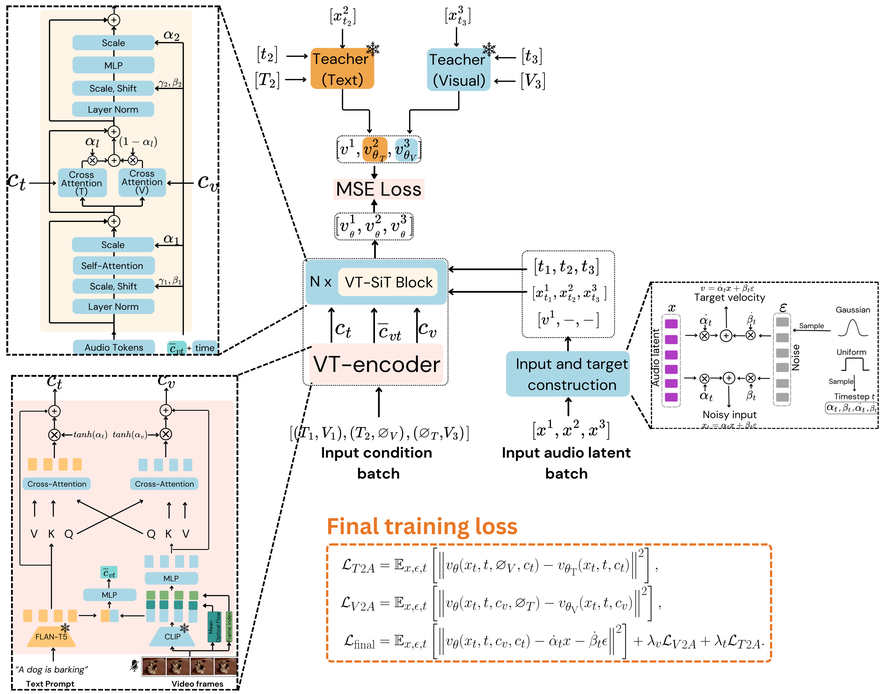


VinTAGe-Bench: Evaluating holistic audio generation

- Subset from VGGSound Test
- *636 datapoints*
- *14 on-screen classes*
- *24 off-screen classes*



Approach: Teacher guidance reduce modality bias



Final training loss

$$\mathcal{L}_{T2A} = \mathbb{E}_{x,\epsilon,t}\left[\left\|v_\theta(x_t,t,\varnothing_V,c_t) - v_{\theta_T}(x_t,t,c_t)\right\|^2\right],$$

$$\mathcal{L}_{V2A} = \mathbb{E}_{x,\epsilon,t}\left[\left\|v_\theta(x_t,t,c_v,\varnothing_T) - v_{\theta_V}(x_t,t,c_v)\right\|^2\right],$$

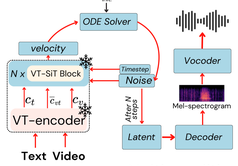
$$\mathcal{L}_{\text{final}} = \mathbb{E}_{x,\epsilon,t}\left[\left\|v_\theta(x_t,t,c_v,c_t) - \dot{\alpha}_t x - \dot{\beta}_t \epsilon\right\|^2\right] + \lambda_v \mathcal{L}_{V2A} + \lambda_t \mathcal{L}_{T2A}.$$

Augmentation and Inference

- Training data : VGGSound

$$\begin{matrix}(v_1,t_1,a_1)\\(v_2,t_2,a_2)\end{matrix} \longrightarrow \begin{matrix}(v_1,t_1+t_2,\text{mix}(a_1,a_2)),\\(v_2,t_1+t_2,\text{mix}(a_1,a_2)),\\(v_1,t_2,\text{mix}(a_1,a_2)),\\(v_2,t_1,\text{mix}(a_1,a_2)).\end{matrix}$$

Augmentations allow creating offscreen sounds



$$\tilde{v}_\theta(x_t,c_v,c_t) = v_\theta(x_t,\varnothing_V,\varnothing_T) + s_{\text{vis}}\cdot(v_\theta(x_t,c_v,\varnothing_T) - v_\theta(x_t,\varnothing_V,\varnothing_T)) + s_{\text{txt}}\cdot(v_\theta(x_t,c_v,c_t) - v_\theta(x_t,c_v,\varnothing_T)),$$

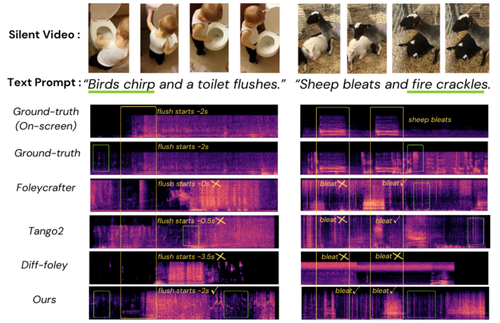
Multi-conditional inference

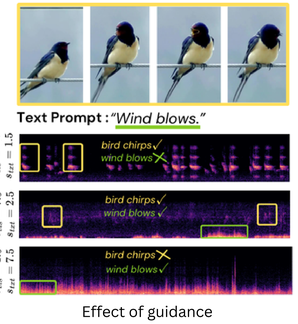
Results: joint guidance improves audio quality

Model	Txt	Vis	Generation Quality FAD↓	FID↓	MKL↓	Alignment AT↑	AV↑	Mean↑	Concept Accuracy(%) On-acc↑	Off-acc↑	Mean↑	Subjective Metrics MOS-Q↑	MOS-F↑	MOS-T↑
SpecVQGAN	✗	✓	6.10	31.73	6.44	18.73	4.61	11.67	25.0	10.61	17.80	1.36	1.08	1.34
Seeing-and-Hearing	✗	✓	5.08	27.04	6.47	15.78	7.93	11.85	34.9	9.43	22.16	-	-	-
Diff-Foley	✗	✓	6.63	19.78	6.38	18.25	9.39	13.82	40.72	7.05	23.88	1.86	1.26	2.36
Make-an-Audio	✓	✗	**4.05**	**19.01**	5.19	18.54	7.42	12.98	52.04	35.14	47.83	-	-	-
AudioLDM2	✓	✗	5.40	20.75	5.52	22.02	6.65	14.33	54.4	26.65	40.52	2.42	2.52	2.10
Tango2	✓	✗	5.85	36.01	4.94	23.84	7.19	15.51	62.57	48.58	**55.57**	2.86	3.34	2.56
Seeing-and-Hearing-VT	✓	✓	4.89	22.44	5.00	17.75	8.99	13.37	43.23	17.68	30.45	-	-	-
Tango2 + LLaVA	✓	✓	4.12	29.1	**4.59**	24.14	7.35	15.75	57.86	40.33	49.09	2.88	**2.90**	2.52
ReWaS	✓	✓	8.01	36.88	7.54	21.03	4.48	12.75	28.14	11.32	19.73	-	-	-
FoleyCrafter	✓	✓	5.81	25.64	4.94	21.36	10.57	**15.96**	64.93	21.69	43.31	**2.92**	2.60	**2.96**
VinTAGe (Ours)	✓	✓	**3.05**	**16.43**	**4.74**	22.29	9.83	**16.06**	57.7	43.63	**50.66**	**3.36**	**3.58**	**3.36**

Comparison on VinTAGe-Bench

Qualitative results





Effect of guidance

Conclusion

- Explore an under-explored **text + video to audio** generation task
- Introduce a **new evaluation benchmark** i.e. VinTAGe-Bench
- New approach and **teacher guidance** to mitigate modality bias

References and Acknowledgement

[1] Luo et. al, "DIFF-FOLEY: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models", NeurIPS, 2023.
[2] Peebles et. al, "Scalable Diffusion Models with Transformers", ICCV, 2023
[3] Ma et. al, "Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers (SiT)", ECCV 2024
[4] Chen et. al, 'VGGSound: A Large-scale Audio-Visual Dataset", ICASSP 2020

Acknowledgement : This work was supported by an Amazon Research Award Fall 2023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of Amazon.