Frieren:

Ffficient Video-to-Audio Gene**r**ati**i**on N**e**twork with **Re**ctified Flow Matchi**n**g

Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, Zhou Zhao
Zhejiang University

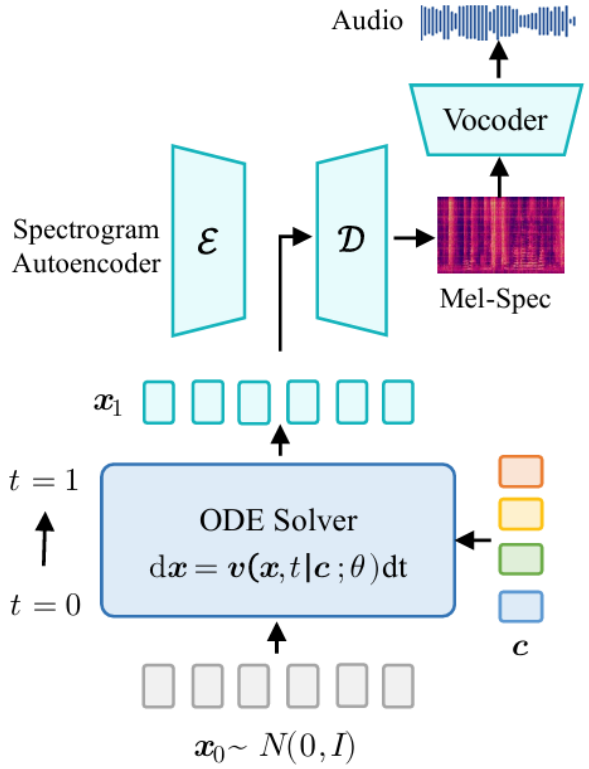
Background: Video-to-Audio (V2A) Generation

- Generate semantically relevant and temporally aligned audio from video frames.
- **Task Focus of V2A**
  - **Audio quality:** the generated audio should have good perceptual quality
  - **Temporal alignment:** the generated audio should not only match the content but also align temporally with the video frames.
  - **Generation efficiency:** the model should be efficient in terms of generation speed and resource utilization.

Our Contribution

- Combining rectified flow matching and feed-forward transformer vector field estimator for higher quality.
- Using channel-wise cross-modal feature fusion for better temporal alignment with simple design.
- Combining reflow and one-step distillation for higher generation efficiency.

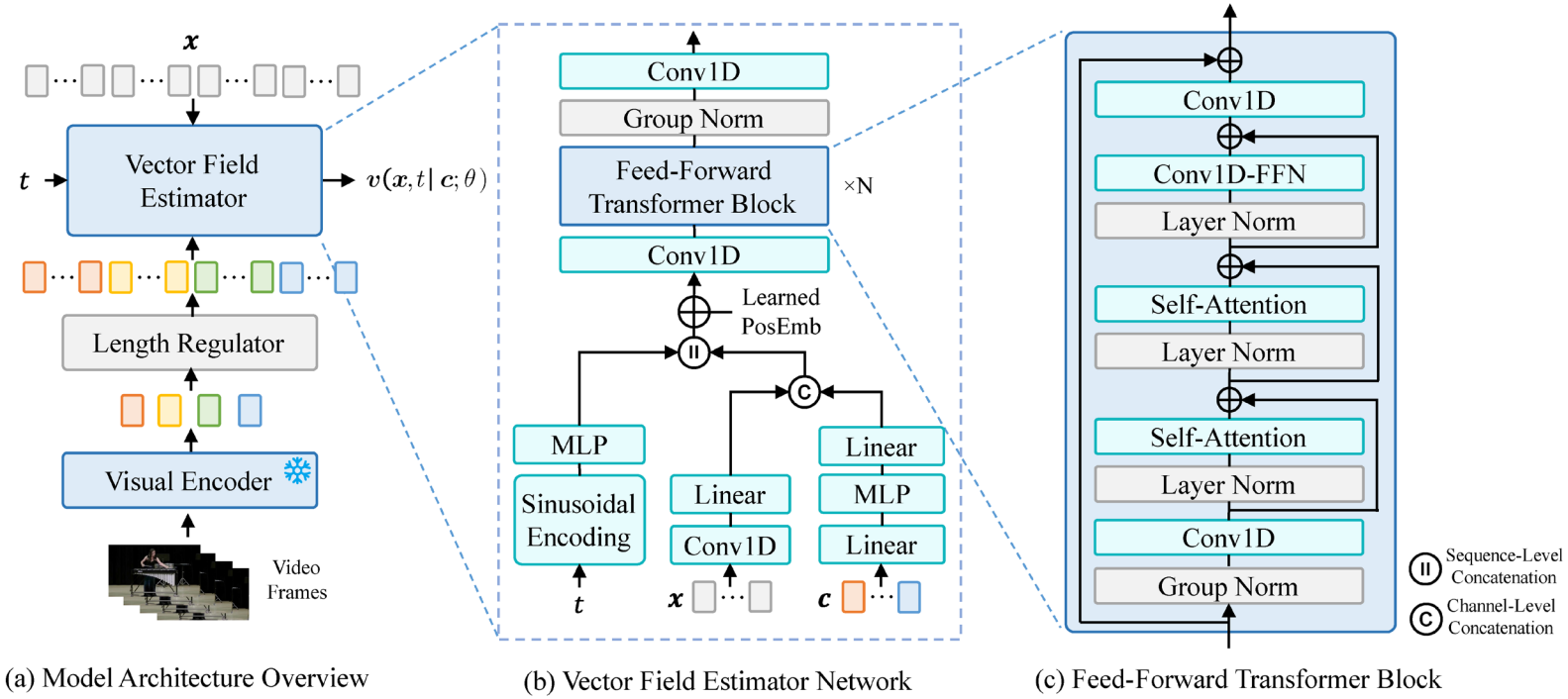
Basic Principle



ODE: $\mathrm{d}\boldsymbol{x} = \boldsymbol{u}(\boldsymbol{x}, t \mid \boldsymbol{c})\mathrm{d}t, t \in [0,1]$

Rectified flow matching objective: $\|\boldsymbol{v}(\boldsymbol{x}, t \mid c; \theta) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|^2$

Model Architecture



Here we use the network in the figure to sample compressed VAE latent of spectrogram conditioned on visual features.

The visual features and point x are processed by some shallow layers and then concatenated along the channel dimension to realize cross-modal feature fusion, achieving good temporal alignment.

CFG, Reflow and Distillation

CFG: $\boldsymbol{v}_{\mathrm{CFG}}(\boldsymbol{x}, t \mid \boldsymbol{c}; \theta) = \gamma \boldsymbol{v}(\boldsymbol{x}, t \mid \boldsymbol{c}; \theta) + (1-\gamma)\boldsymbol{v}(\boldsymbol{x}, t \mid \emptyset; \theta)$

Reflow objective: $\mathcal{L}_{\mathbf{reflow}}(\theta') = \mathbb{E}_{t, p(x_0', \hat{\boldsymbol{x}}_1 \mid c), p_t(\boldsymbol{x} \mid \boldsymbol{x}_0', \hat{\boldsymbol{x}}_1)} \| \boldsymbol{v}_{\mathrm{CFG}}(\boldsymbol{x}, t \mid c; \theta') - (\hat{\boldsymbol{x}}_1 - \boldsymbol{x}_0') \|^2$

Distillation objective: $\mathcal{L}_{\mathbf{distill}}(\theta'') = \mathbb{E}_{t, p(\boldsymbol{x}_0', \hat{\boldsymbol{x}}_1 \mid c), p_t(\boldsymbol{x} \mid \boldsymbol{x}_0', \hat{\boldsymbol{x}}_1)} \| \boldsymbol{x}_0' + \boldsymbol{v}_{\mathrm{CFG}}(\boldsymbol{x}_0', t \mid \boldsymbol{c}; \theta'') - \hat{\boldsymbol{x}}_1 \|^2$

Results

Superior performance across multiple metrics

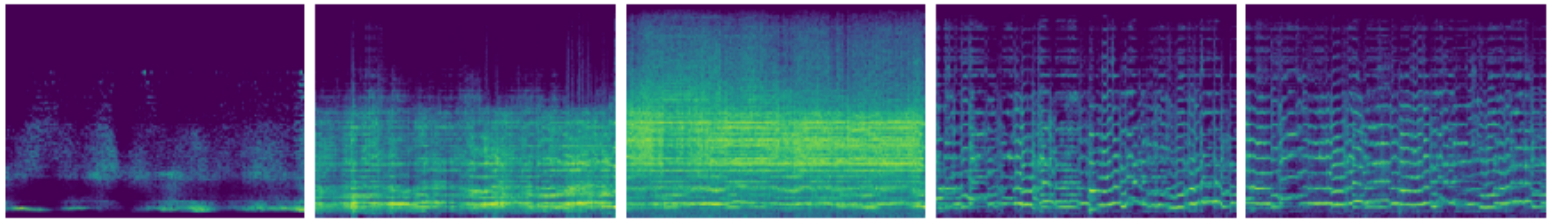
Model	FD↓	IS↑	KL↓	FAD↓	KID(10^{-3}) ↓	Acc(%) ↑	MOS-Q↑	MOS-A↑
SpecVQGAN (R+F)	31.69	5.23	3.37	5.42	8.53	61.83	3.30 ± 0.06	2.35 ± 0.05
SpecVQGAN (RN50)	32.52	5.21	3.41	5.39	9.00	56.92	3.25 ± 0.07	2.17 ± 0.05
Im2Wav	14.98	7.20	**2.57**	5.49	3.35	56.70	3.39 ± 0.06	2.29 ± 0.06
Diff-Foley (CG ✓)	23.94	11.11	3.38	4.72	9.58	95.03	3.57 ± 0.08	3.74 ± 0.07
Diff-Foley (CG ✗)	24.97	11.69	3.23	7.10	10.32	92.53	3.64 ± 0.07	3.59 ± 0.06
LDM	11.79	10.09	2.86	1.77	**2.36**	95.33	3.72 ± 0.05	3.79 ± 0.07
FRIEREN	12.26	12.42	2.73	**1.32**	2.49	**97.22**	3.78 ± 0.06	$\mathbf{3.90 \pm 0.05}$
FRIEREN (Dopri5)	**11.64**	**12.76**	2.75	1.37	2.39	96.87	$\mathbf{3.81 \pm 0.06}$	3.85 ± 0.06

Results

Good few-step and one-step performance with reflow and distillation

Model	Steps	FD↓	IS↑	KL↓	FAD↓	KID(10^{-3}) ↓	Acc(%) ↑	MOS-Q↑	MOS-A↑
Diff-Foley (CG ✓)	1	82.61	2.31	4.44	13.64	43.96	31.60	1.28 ± 0.04	1.35 ± 0.03
Diff-Foley (CG ✗)		86.97	1.86	4.17	14.66	39.73	37.02	1.17 ± 0.03	1.63 ± 0.04
FRIEREN (R ✗, D ✗)		70.48	2.95	4.21	13.07	26.99	43.18	2.12 ± 0.04	1.71 ± 0.04
FRIEREN (R ✓, D ✗)		18.61	6.63	2.60	3.13	3.49	94.96	3.32 ± 0.07	3.74 ± 0.06
FRIEREN (R ✓, D ✓)		**17.58**	**8.66**	**2.56**	**1.85**	**2.91**	**97.85**	**3.48 ± 0.06**	**3.93 ± 0.05**
Diff-Foley (CG ✓)	5	60.99	3.42	3.62	9.61	3.60	73.30	2.66 ± 0.07	2.98 ± 0.07
Diff-Foley (CG ✗)		51.52	5.14	3.45	10.96	2.66	91.30	3.03 ± 0.08	3.56 ± 0.07
FRIEREN (R ✗, D ✗)		28.78	6.69	3.02	4.34	8.56	87.69	3.30 ± 0.07	3.37 ± 0.08
FRIEREN (R ✓, D ✗)		**14.65**	**8.28**	**2.60**	**2.11**	**2.28**	**96.82**	**3.43 ± 0.06**	**3.83 ± 0.06**
Diff-Foley (CG ✓)	25	23.94	11.11	3.28	4.72	9.58	95.03	3.57 ± 0.08	3.74 ± 0.07
Diff-Foley (CG ✗)		24.97	11.69	3.23	7.10	10.32	92.53	3.64 ± 0.07	3.59 ± 0.06
FRIEREN (R ✗, D ✗)		12.26	**12.42**	2.73	**1.32**	2.49	97.22	**3.78 ± 0.06**	**3.90 ± 0.05**
FRIEREN (R ✓, D ✗)		13.39	9.79	**2.64**	1.66	2.01	**97.36**	3.61 ± 0.07	3.88 ± 0.05

(a) Input video



(b) Diff-Foley (c) LDM (d) FRIEREN (no reflow) (e) FRIEREN (reflow) (f) FRIEREN (reflow + distillation)

Results

High time efficiency

Model	Inference Time (sec)
SpecVQGAN	3.936
Im2Wav	333.246
Diff-Foley (step=25)	2.104
FRIEREN (Dopri5, step=25)	1.510
FRIEREN (Euler, step=25)	0.288
FRIEREN (Euler, step=5)	0.064
FRIEREN (Euler, step=1)	0.031

Thanks for Listening.