Few-shot Acoustic Synthesis with Multimodal Flow Matching

Amandine Brunetto
Mines Paris - PSL University

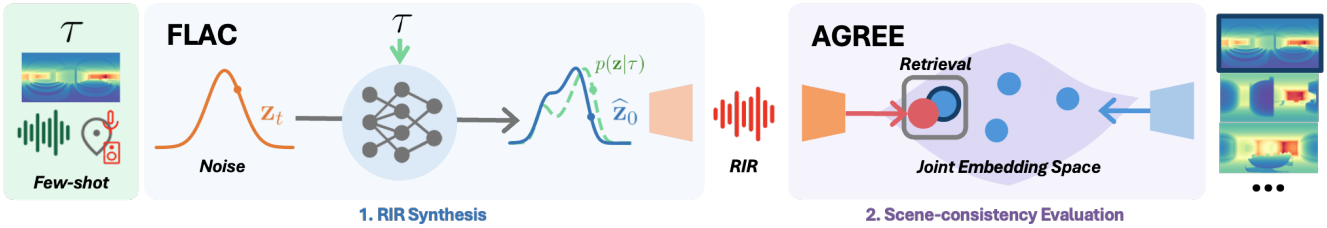

Figure 1. **Few-shot flow-matching acoustic synthesis (FLAC):** Given few-shot multimodal context τ (depth map, acoustic observations, sensor poses), FLAC, a diffusion transformer trained with flow matching, generates RIRs in novel rooms. It captures the inherent ambiguity of the task by modeling the distribution of plausible RIRs under sparse conditioning. With a single shot, FLAC outperforms 8-shot state-of-the-art methods. We introduce AGREE, a CLIP-style audio-geometry embedding enabling scene-consistency evaluation.

Abstract

Generating audio that is acoustically consistent with a scene is essential for immersive virtual environments. Few-shot acoustic synthesis approaches improve scalability across rooms but are deterministic, failing to capture the inherent uncertainty of scene acoustics under sparse context. We introduce flow-matching acoustic generation (FLAC), a probabilistic method for few-shot acoustic synthesis that models the distribution of plausible room impulse responses (RIRs) given minimal scene context. FLAC leverages a diffusion transformer trained with a flow-matching objective to generate RIRs at arbitrary positions in novel scenes, conditioned on spatial, geometric, and acoustic cues. FLAC outperforms state-of-the-art eight-shot baselines with one-shot on both the AcousticRooms and Hearing Anything Anywhere datasets. We further introduce AGREE, a joint acoustic-geometry embedding, enabling geometry-consistent evaluation of generated RIRs. Project page: https://amandinebtto.github.io/FLAC/

1. Introduction

Every room shapes the way we hear: a lecture hall amplifies a speaker's voice, while a cathedral envelops sound in lingering reverberation. Reproducing these rich auditory experiences is essential for creating virtual, immersive environments, where users expect sound to reflect the space. The acoustic properties of a room are encapsulated by Room Impulse Responses (RIRs), which describe sound propagation between source-receiver pairs and enable auralization. Accurately modeling RIRs is challenging because they depend on complex interactions between geometry, materials, and sensor positions. Neural acoustic fields [2, 3, 12, 16, 26, 27] achieve spatially continuous RIR rendering but require per-scene training with dense recordings. Few-shot acoustic synthesis [9, 15, 18] addresses this by generating RIRs from sparse observations. With limited scene information, multiple RIRs can be equally plausible, *e.g.*, unknown floor material significantly changes acoustics, making few-shot synthesis an inherently ambiguous problem, yet existing methods overlook this uncertainty.

We propose FLAC, a conditional latent generative flow matching [14] model for few-shot acoustic synthesis. Rather than learning a deterministic mapping, FLAC estimates a distribution of plausible RIRs given sparse scene context, conditioned on scene geometry, sensor poses, and a minimal set of RIR recordings. To our knowledge it is the first application of generative flow matching to explicit RIR synthesis. We further introduce AGREE (Acoustic-GeometRy EmbEding), a CLIP-style [20] joint embedding space for RIRs and scene geometry, enabling geometry-consistent evaluation via retrieval and distributional metrics. Our main contributions are:

- FLAC, the first conditional generative model for few-shot RIR synthesis based on flow matching, capturing acoustic uncertainty under sparse context.
- FLAC one-shot outperforms eight-shot state-of-the-art baselines on AcousticRooms and Hearing-Anything-Anywhere datasets.
- AGREE, a joint acoustic-geometry embedding enabling new scene-consistency evaluation metrics.

2. Related Work

Neural acoustic fields [2, 3, 12, 16, 26, 27] render RIRs at novel poses by implicitly learning a mapping from spatial coordinates to the room's acoustic field, but require

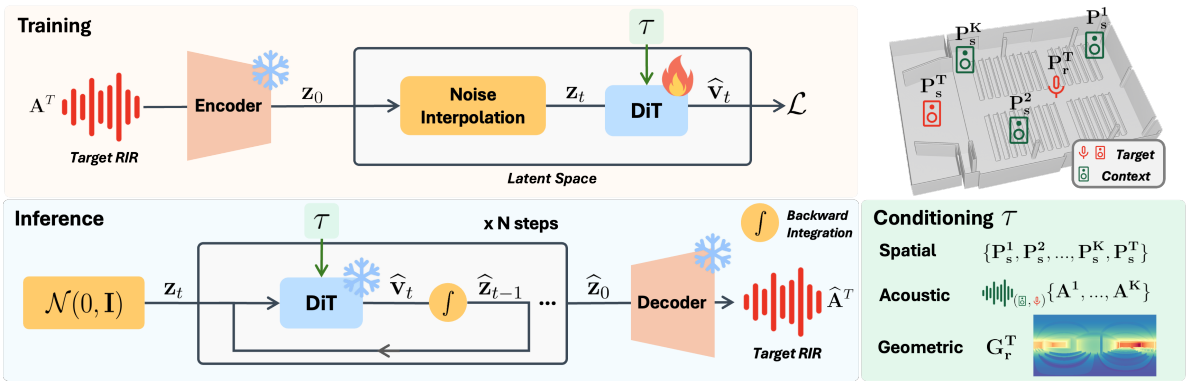

Figure 2. **Training and inference pipelines of FLAC:** A VAE encodes ground-truth RIRs into latents $\mathbf{z}_0$, linearly interpolated with noise to form $\mathbf{z}_t$. A DiT predicts the velocity $\hat{\mathbf{v}}_t$. At inference, RIRs are generated from noise guided by few-shot multimodal context.

per-scene training with dense recordings. Few-shot methods [9, 15, 18] generalize across scenes using sparse observations (e.g., RGB, depth, 8-20 RIRs) but produce a single deterministic output, ignoring the inherent ambiguity of the task. FLAC addresses this with stochastic generative modeling, adapting flow matching [14], previously applied to speech and music [10, 11, 13], to explicit RIR synthesis. For evaluation, we build AGREE inspired by CLIP-style audio-visual representation learning [5, 6, 17, 19, 22].

3. Method

FLAC is a conditional latent generative model [23] trained with flow matching [1, 14] to predict monaural, omnidirectional RIRs at arbitrary source-receiver pairs in unseen environments, given minimal scene context. It synthesizes RIRs conditioned on few-shot scene context and comprises: (i) a variational autoencoder (VAE), (ii) a multimodal conditioner, and (iii) a diffusion transformer (DiT). Fig. 2 provides an overview.

Latent Flow Matching. At training time, we sample target RIRs with their associated context $(A^T, \boldsymbol{\tau})$ from the dataset. Each RIR is encoded into a latent representation $\mathbf{z}_0$, which is linearly interpolated with noise ϵ to produce $\mathbf{z}_t$:

$$\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where the timestep $t \in [0,1]$ controls the noise level. The model $u(\mathbf{z}_t, t, \boldsymbol{\tau})$ predicts velocity

$$\mathbf{v}_t = \frac{d\mathbf{z}_t}{dt} = \boldsymbol{\epsilon} - \mathbf{z}_0, \tag{2}$$

using the following objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{z}_0, \boldsymbol{\epsilon}, t, \boldsymbol{\tau}}\Big[\, \| u(\mathbf{z}_t, t, \boldsymbol{\tau}) - \mathbf{v}_t \|^2 \Big]. \tag{3}$$

At inference, we use classifier-free guidance [8] to obtain $\hat{u}(\mathbf{z}_t, t, \boldsymbol{\tau})$ and generate RIRs by solving the ODE backward from $t = 1$ to $t = 0$:

$$\mathbf{z}_{t-dt} = \mathbf{z}_t + \hat{u}(\mathbf{z}_t, t, \boldsymbol{\tau})\, dt. \tag{4}$$

Table 1. **Performance on unseen AR scenes:** Results are shown for $K \in \{8, 1, \text{✗}\}$ reference RIRs. For FLAC, we report mean and standard deviation over 5 generations. FLAC outperforms all baselines even in the one-shot setting. ✖ denotes modality ablations.

Method	K	G	T60 (%) ↓	C50 (dB) ↓	EDT (ms) ↓	R@5 (%) ↑	FD_G ↓
Random Across Rooms	✗	✗	44.73	7.676	306.29	0.06	0.111
Random Same Room	✗	✗	**17.36**	5.490	168.17	1.09	**0.001**
FLAC✖	✗	✓	23.41$_{\pm 0.02}$	**2.554**$_{\pm 0.002}$	**109.75**$_{\pm 0.09}$	**16.47**$_{\pm 0.14}$	0.337
Nearest Neighbor	1	✗	15.22	5.212	157.94	2.26	**0.001**
Fast-RIR	1	✓	18.97	3.257	121.21	0.66	0.456
xRIR	1	✓	14.47	1.961	74.45	1.36	0.263
FLAC	1	✓	**9.95**$_{\pm 0.05}$	**1.046**$_{\pm 0.002}$	**40.04**$_{\pm 0.22}$	**18.92**$_{\pm 0.10}$	0.303
Linear Interpolation	8	✗	14.45	3.503	114.27	2.30	0.401
Nearest Neighbor	8	✗	10.91	2.792	90.08	10.26	**0.003**
FLAC✖	8	✗	12.07$_{\pm 0.01}$	4.296$_{\pm 0.001}$	140.04$_{\pm 0.04}$	0.58$_{\pm 0.06}$	0.663
Fast-RIR	8	✓	17.71	3.253	121.21	0.99	0.465
xRIR	8	✓	9.98	1.354	49.40	2.00	0.307
FLAC	8	✓	**8.60**$_{\pm 0.01}$	**0.970**$_{\pm 0.002}$	**37.13**$_{\pm 0.02}$	**19.38**$_{\pm 0.15}$	0.305

Table 2. **Sim-to-real transfer to the HAA dataset:** Few-shot methods are compared against Diff-RIR and INRAS, which require per-scene training (†).

Method	K	T60 (%) ↓	C50 (dB) ↓	EDT (ms) ↓	R@5 (%) ↑	FD_G ↓
Random Across Rooms	✗	17.40	10.283	533.99	1.49	0.460
Random Same Room	✗	8.00	4.805	180.15	1.86	0.169
Nearest Neighbor	1	8.19	5.000	187.55	1.20	**0.177**
xRIR	1	8.63	4.862	183.27	14.85	0.363
FLAC	1	**3.45**$_{\pm 0.02}$	**2.170**$_{\pm 0.014}$	**90.02**$_{\pm 0.24}$	**17.94**$_{\pm 0.62}$	0.564
Linear Interpolation	8	4.12	2.695	88.19	3.62	0.904
Nearest Neighbor	8	**2.89**	**1.923**	**77.24**	9.61	**0.169**
xRIR	8	6.53	3.492	149.69	**20.65**	0.318
FLAC	8	3.10$_{\pm 0.01}$	2.167$_{\pm 0.004}$	84.52$_{\pm 0.24}$	17.41$_{\pm 0.59}$	0.585
INRAS†	12	6.61	3.966	158.07	2.27	0.797
Diff-RIR†	12	3.74	2.067	88.09	26.97	0.263

VAE. We found pre-trained audio embedding proved unsuitable for RIR synthesis, so we train a VAE to compress RIR waveforms into compact latents $\mathbf{z}_0$. The encoder uses four strided convolutional blocks with ResNet-style dilated layers and Snake activations [28], while the decoder mirrors this design; the bottleneck has dimension 32. To preserve fine temporal and spectral structure, we optimize a combination of multiresolution STFT, adversarial and feature-matching losses (with an Encodec [4] discriminator), and a KL divergence term.

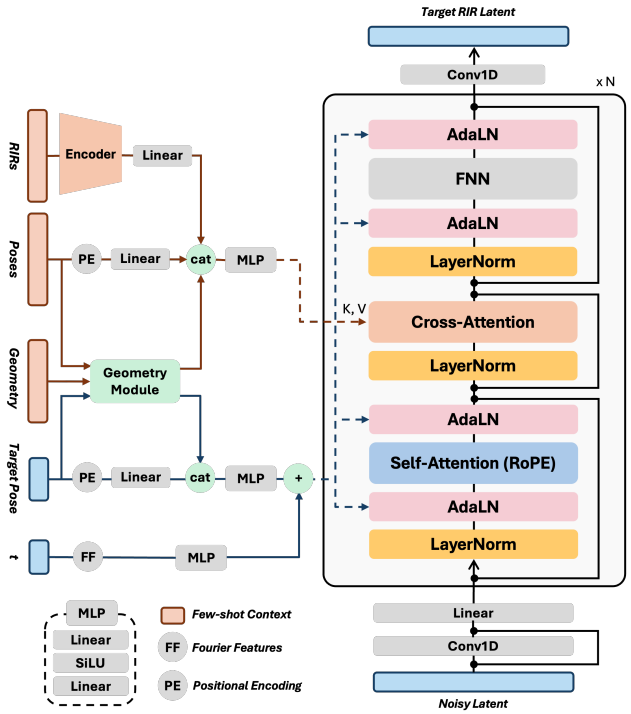

Figure 3. **FLAC DiT:** Timestep t and target pose are injected via AdaLN; multimodal context via cross-attention.

Multimodal Conditioning. FLAC generates RIRs at target (P_s^T, P_r^T) conditioned on: **(Acoustic)** K RIR recordings, encoded via ResNet-18; **(Spatial)** source poses in the receiver frame, encoded with sinusoidal positional embeddings; **(Geometric)** a panoramic depth map captured at the receiver position is converted to reflection maps and encoded via fine-tuned DINOv3 ViT-S/16 [24].

Diffusion Transformer. The DiT (Fig. 3) consists of transformer blocks with RoPE [25] self-attention, cross-attention, and a feedforward network. The target pose and timestep are injected via AdaLN and multimodal context via cross-attention.

4. AGREE: Acoustic-Geometry Embedding

AGREE is a joint embedding that aligns room acoustics and geometry. The audio encoder is the pre-trained FLAC VAE fine-tuned. The geometry encoder is DINOv3 ViT-S/16 fine-tuned on reflection maps obtained from the panoramic depth maps. Each encoder is followed by a linear projection, and trained jointly with a contrastive objective. We use AGREE to define two scene-consistency metrics: **audio-to-audio recall** (alignment of generated vs. ground-truth RIRs in geometry-aware space) and **Fréchet distance** FD_G (distributional realism, analogous to FID [7]).

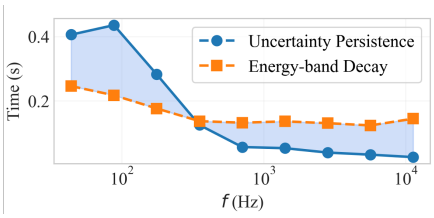

Figure 4. **Uncertainty persistence time** and band-wise energy decay, averaged over 100 unseen samples. Uncertainty lasts longer at low frequencies and decays faster at high frequencies.

5. Results

Experimental Setup. The AcousticRooms (AR) [15] dataset contains over 300k simulated monaural RIRs across 260 rooms. Cross-room generalization is assessed on 17 held-out, unseen rooms (5,244 samples). For sim-to-real transfer, we use Hearing Anything Anywhere (HAA) [27], which provides real-world RIRs from four rooms. We report perceptual metrics (T60, C50, EDT errors) and scene-consistency metrics (R@5 and FD_G in AGREE space). Baselines include non-learning methods (Random, KNN, Linear Interpolation) and learning-based approaches (FastRIR [21], xRIR [15]).

Few-shot synthesis in unseen rooms. Quantitative results are reported in Tab. 1. With K=8, FLAC outperforms xRIR: -13.8% T60, -28.3% C50, -24.9% EDT. FLAC remains significantly more stable than KNN and xRIR as K decreases. Crucially, **1-shot FLAC surpasses all 8-shot baselines.** Geometry conditioning provides the dominant cue for scene-consistent synthesis.

Sim-to-Real Transfer. Following [15], we fine-tune xRIR and FLAC (except VAE). Results are in Tab. 2. With K=8, FLAC outperforms xRIR. It also surpasses DiffRIR [27] and INRAS [26], which require per-scene training, on most perceptual metrics. In the 1-shot setting, FLAC outperforms both KNN and 8-shot xRIR.

Capturing uncertainty. When generating 100 RIRs for each conditioning, uncertainty persistence time, defined as the time until band-wise sample variance drops below the 75th percentile, is longer at low frequencies (Fig. 4) This is physically grounded: below the Schroeder frequency, responses are governed by sparse boundary-dependent modes weakly constrained by limited context, while denser high-frequency modes are stabilized by local geometry. A deterministic variant degrades performance ($+6\%$ T60, $+10\%$ C50, -40% R@5), confirming stochasticity is essential.

Perceptual evaluation. A listening study with 46 participants on 14 unseen AR scenes was conducted. Participants were presented with the ground-truth, audio generated by FLAC (1-shot) and xRIR (8-shot), and were asked to select which audio sounded closer to the GT. FLAC was preferred in 93% of cases.

6. Conclusion

We introduced FLAC, a generative flow-matching approach for few-shot acoustic synthesis, and AGREE, a joint acoustic-geometry embedding for scene-consistent evaluation. FLAC captures the ambiguity of few-shot RIR synthesis overlooked by prior deterministic methods, achieving state-of-the-art performance with as few as one reference recording. Future work includes supporting multiple sample rates and collecting larger real-world audio-visual data for improved sim-to-real transfer. AGREE embedding could also benefit broader audio-visual learning tasks.

References

[1] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023. 2

[2] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. AV-GS: Learning material and geometry aware priors for novel view acoustic synthesis. In *NeurIPS*, 2024. 1

[3] Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. NeRAF: 3D scene infused neural radiance and acoustic fields. In *ICLR*, 2025. 1

[4] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. 2

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 2

[6] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, 2022. 2

[7] 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. *NeurIPS*, 2017. 3

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPSW*, 2021. 2

[9] Diwei Huang, Kunyang Lin, Peihao Chen, and Qing Du. Map-guided few-shot audio-visual acoustics modeling. In *ICASSP*, 2025. 1, 2

[10] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024. 2

[11] Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. ETTA: Elucidating the design space of text-to-audio models. In *ICML*, 2025. 2

[12] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In *NeurIPS*, 2023. 1

[13] Susan Liang, Dejan Markovic, Israel D Gebru, Steven Krenn, Todd Keebler, Jacob Sandakly, Frank Yu, Samuel Hassel, Chenliang Xu, and Alexander Richard. Binauralflow: A causal and streamable approach for high-quality binaural speech synthesis with flow matching models. In *ICML*, 2025. 2

[14] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 1, 2

[15] Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastià V. Amengual Garí, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *CVPR*, 2025. 1, 2, 3

[16] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *NeurIPS*, 2022. 1

[17] Tanvir Mahmud, Shentong Mo, Yapeng Tian, and Diana Marculescu. Ma-avt: Modality alignment for parameter-efficient audio-visual transformers. In *CVPR*, 2024. 2

[18] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *NeurIPS*, 2022. 1, 2

[19] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. 2

[20] 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*, 2021. 1

[21] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. In *ICASSP*, 2022. 3

[22] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *CVPR*, 2024. 2

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

[24] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3

[25] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 3

[26] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. In *NeurIPS*, 2022. 1, 3

[27] Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. 1, 3

[28] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In *NeurIPS*, 2020. 2