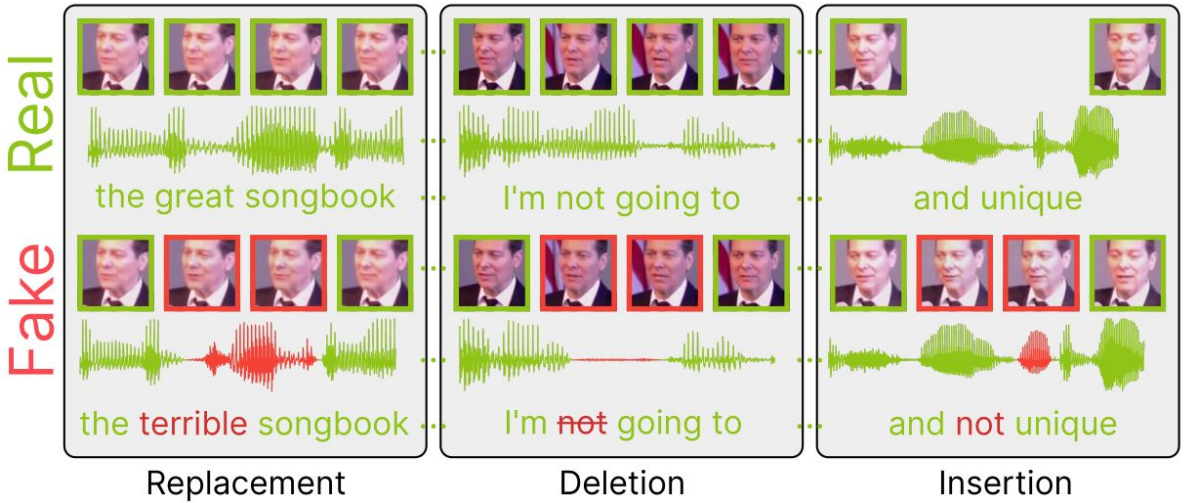
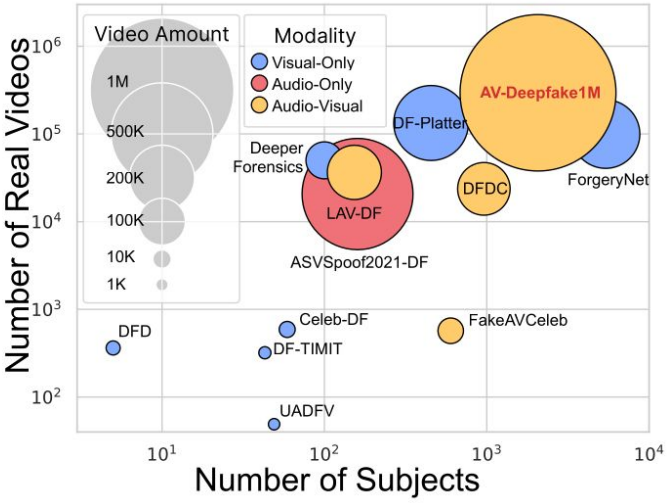
Thematic session

Paper presentation: Group3

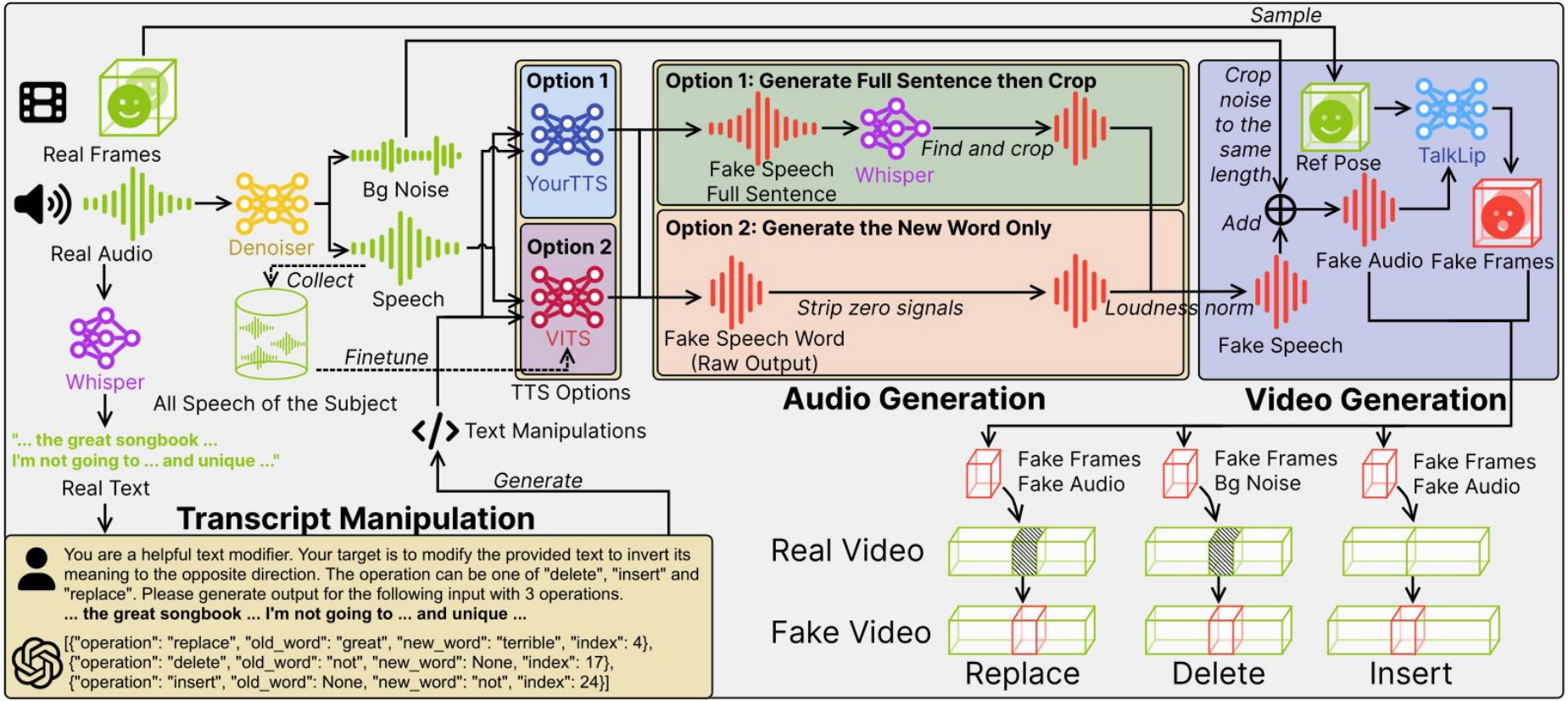
Stevan Rudinac





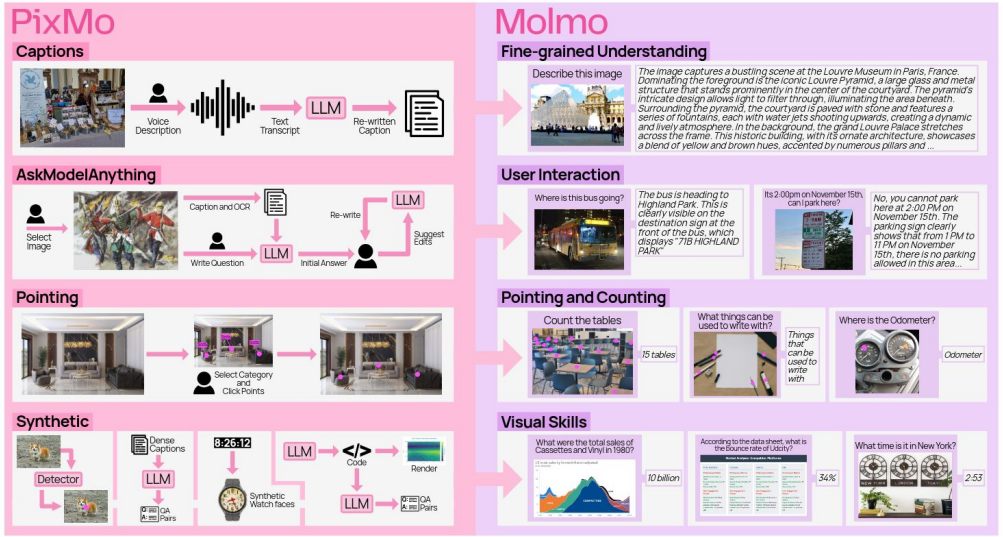
- AV-Deepfake1M is a large-scale content-driven deepfake dataset generated using a large language model.
- Best Student Paper Award at ACM Multimedia 2024 in Melbourne.

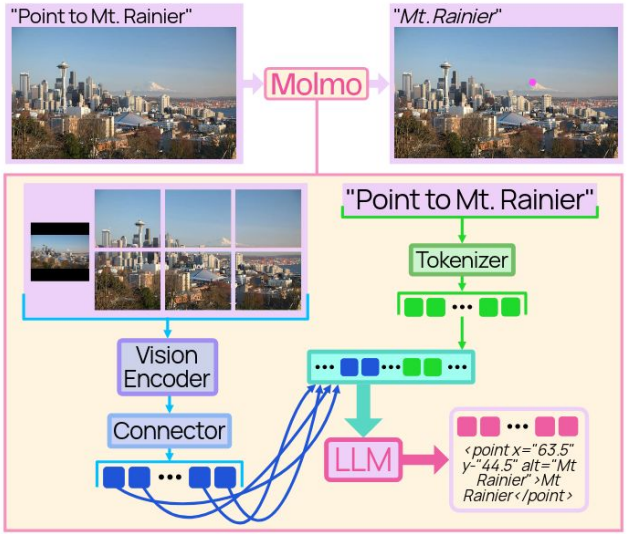
Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. 2024. AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24). Association for Computing Machinery, New York, NY, USA, 7414–7423. https://doi.org/10.1145/3664647.3680795



- **Preprocessing**: Audio extraction via FFmpeg followed by Whisper-based transcript generation.
- **Step 1 (transcript manipulation)**: The transcript is modified through word-level insertions, deletions & replacements.
- **Step 2 (audio generation)**: The audio is generated in both speaker-dependent and independent fashion.
- **Step 3 (video generation)**: Based on the generated audio, the subject-dependant video is generated with smooth transitions in terms of lip-synchronization, pose, and other relevant attributes.

Takeaway from Andrei Bursuc's MMM'25 Keynote





- Careful and smart data selection and annotation can go a long way
- Molmo is a very competitive VLM from Ai2, trained on 700k image/caption pairs
- 3 annotations per image; annotation speech is recorded for 60-90 seconds; formatted questions

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., ... & Kembhavi, A. (2024). Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*. (https://molmo.allenai.org/)

Yen-Chia Hsu

Exploring Empty Spaces: Human-in-the-Loop Data Augmentation

Catherine Yeh*
Harvard University
Boston, Massachusetts, USA
catherineyeh@g.harvard.edu

Donghao Ren
Apple
Seattle, Washington, USA
donghao@apple.com

Yannick Assogba
Apple
Cambridge, Massachusetts, USA
yassogba@apple.com

Dominik Moritz
Apple
Pittsburgh, Pennsylvania, USA
domoritz@apple.com

Fred Hohman
Apple
Seattle, Washington, USA
fredhohman@apple.com

Source – https://dl.acm.org/doi/10.1145/3706598.3713491 (ACM CHI 2025)

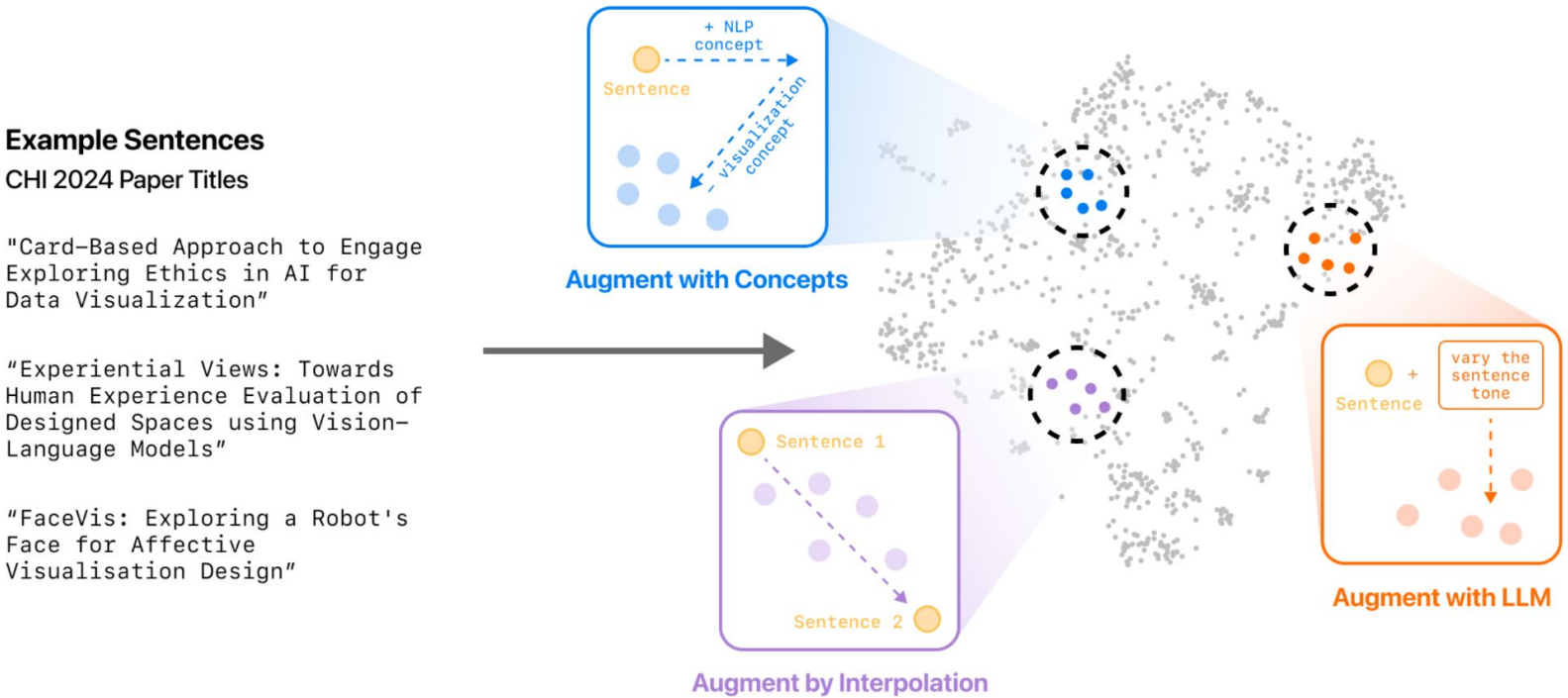


Figure 1: Given a dataset of unstructured text, it can be challenging to determine how and where to augment the data most effectively. We propose a visualization-based approach to help users find relevant *empty data spaces* to explore to improve dataset diversity. To fill in these empty spaces, metaphorically represented by gaps in an embedding plot, we design an interactive tool with three human-in-the-loop augmentation methods: Augment with Concepts, Augment by Interpolation, and Augment with Large Language Model (LLM). Here, each dot represents an embedded sentence from the input dataset of CHI 2024 paper titles [37].

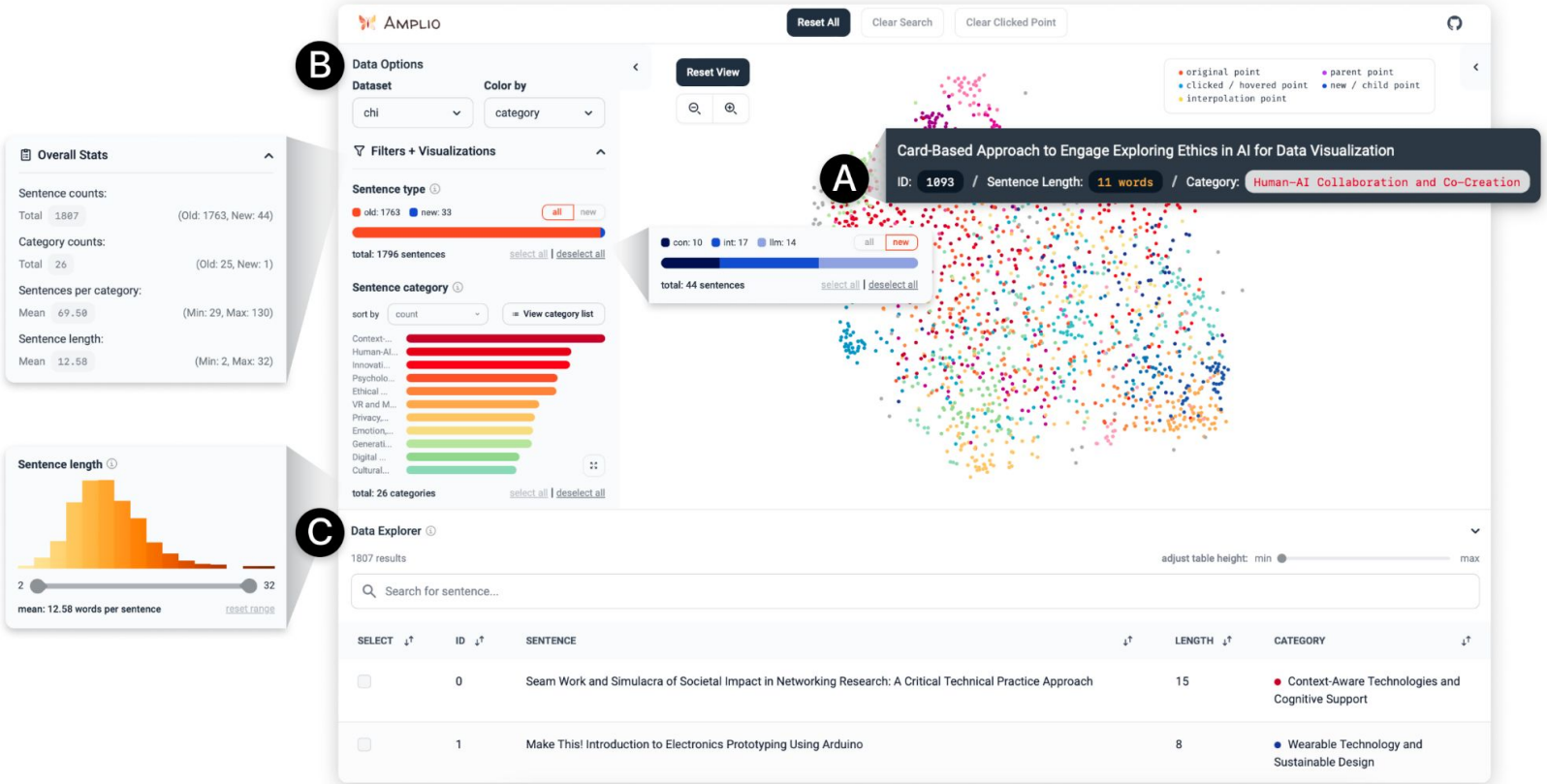


Figure 3: With our interface, ML practitioners can quickly get an overview of their dataset in three ways. (A) First, users can hover over points in the main embedding visualization and view information about the corresponding sentence. (B) The Left Sidebar includes summary statistics and interactive visualizations that can be used to filter the data by sentence type, category, or length. (C) In the Data Explorer view, users can search for specific data instances with a searchable table.

Augment with Concepts

Original: Card-Based Approach to Engage Exploring Ethics in AI for Data Visualization

Concepts: `Ethics, Values, Morality (-0.5); Cardinals, Religious Figures, Sports Teams (+1)`

→ **New Sentence:** Cardinal Cards: An Engaging Card-Based Method for AI-Driven Statistical Data Exploration

Augment by Interpolation

Original: Card-Based Approach to Engage Exploring Ethics in AI for Data Visualization

Interpolation Point: `Footprints of Travel: AIoT and AR Enhanced Tourist Gaming Experience in Unmanned Cultural Sites`

→ **New Sentences:** Card-Based Approach to AI: Exploring Cultural Experiences in the Process of Using Cartography to Visualize Unstructured Data and Ethics ($\alpha = 0.25$).

Guided Travel with AR-AI Experiences and AIoT: Investigating Carded Footprints in Cultural Tourism while Developing Advanced Gaming Solutions ($\alpha = 0.63$).

Augment with LLM

Original: Card-Based Approach to Engage Exploring Ethics in AI for Data Visualization

Prompt: `Create alternative phrases that describe the card-based approach in various contexts related to data visualization`

→ **New Sentence:** Implementing a card-driven framework to examine ethical considerations surrounding AI in the context of data visualization

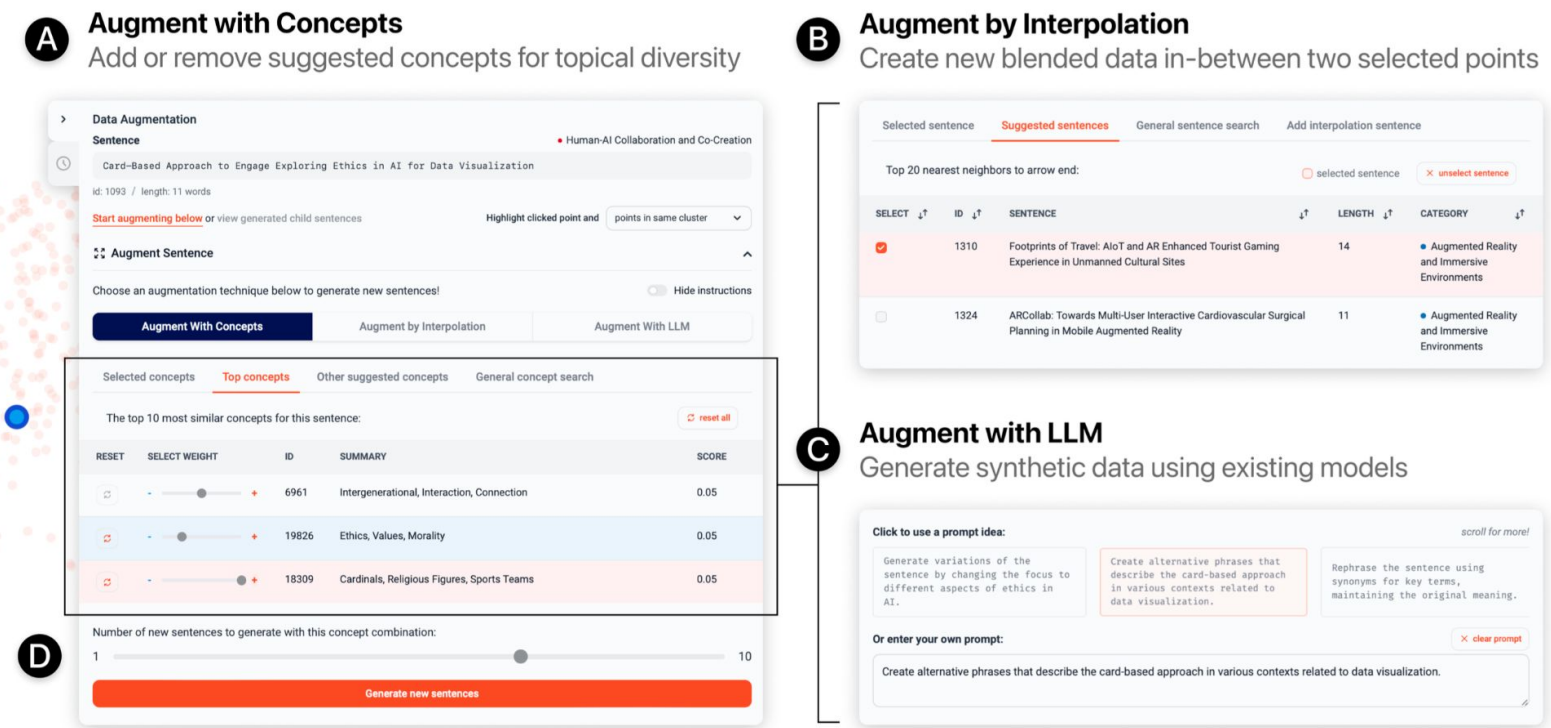
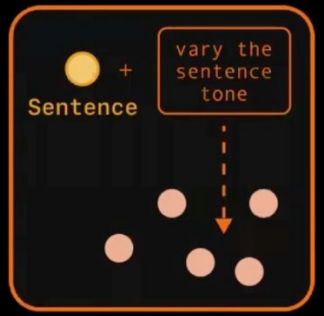


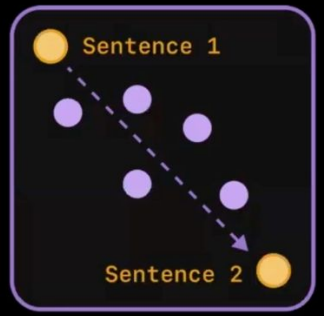
Figure 4: When a user clicks on a point, the data augmentation panel will open on the right. Here, users can choose an augmentation approach. (A) Our first method, *Augment with Concepts* will suggest relevant concepts, which can be added or subtracted from the current sentence by adjusting the weight sliders. (B) Second, to *Augment by Interpolation*, users can select a second sentence to interpolate with to generate new variations. (C) Finally, users can *Augment with Large Language Model* by entering their own prompt, or selecting an prompt idea from the provided list of contextualized suggestions. (D) Below each augmentation method, users can set how many new sentences they would like to generate.

User Study (18 experienced LLM red teamers at Apple)



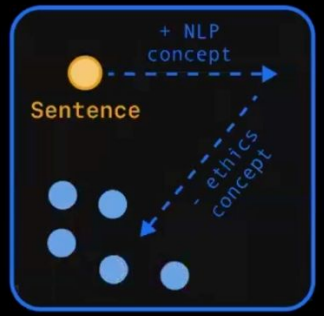
Augment with LLM

"Why is it focusing on knees? I didn't add that... That's an interesting hallucination." (P11)



Augment by Interpolation

*"It didn't get what I wanted. This is just combining them. It doesn't really make a point about credit cards **and** Donald Trump." (P1)*



Augment with Concepts

"I wasn't expecting the prompt ideas to be tailored, so that was super useful." (P9)

- 👍 Most relevant + easy to use
- 👍 Suggested prompt ideas
- 👎 Repetitive outputs

- 👍 Creative sentence blends
- 👎 Abrupt, unnatural interpolations
- 💡 Category interpolation

- 👍 Increased topical diversity
- 👎 Unexpected concepts
- 💡 Custom concepts

Source – https://www.youtube.com/watch?v=6aziBfOL5KM

Tim Alpherts

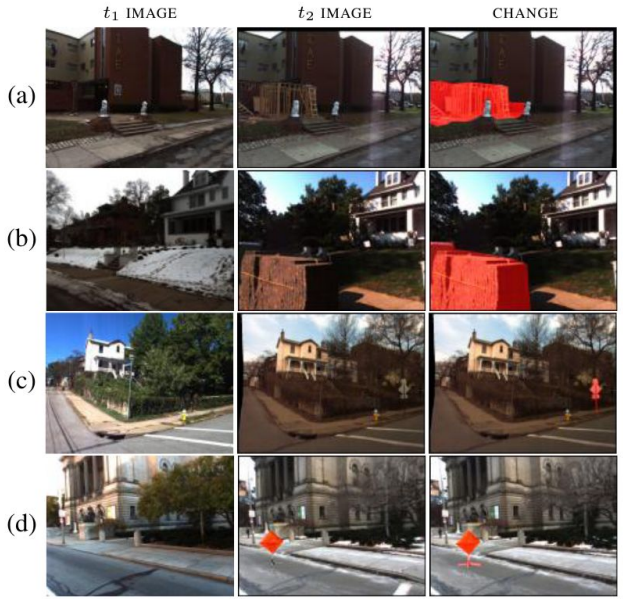
Zero-Shot Scene Change Detection

Kyusik Cho[1], Dong Yeop Kim[2,1], Euntai Kim[1]*

[1]Yonsei University, Seoul, Republic of Korea
[2]Korea Electronics Technology Institute, Seoul, Republic of Korea
ks.cho@yonsei.ac.kr, sword32@keti.re.kr, etkim@yonsei.ac.kr

Problem Statement

- Scene Change Detection (SCD) aims to detect differences between two scenes.
- Research has focused on supervised methods.
- Collecting supervised data is labour intensive.
- Must be season and weather invariant.



Method

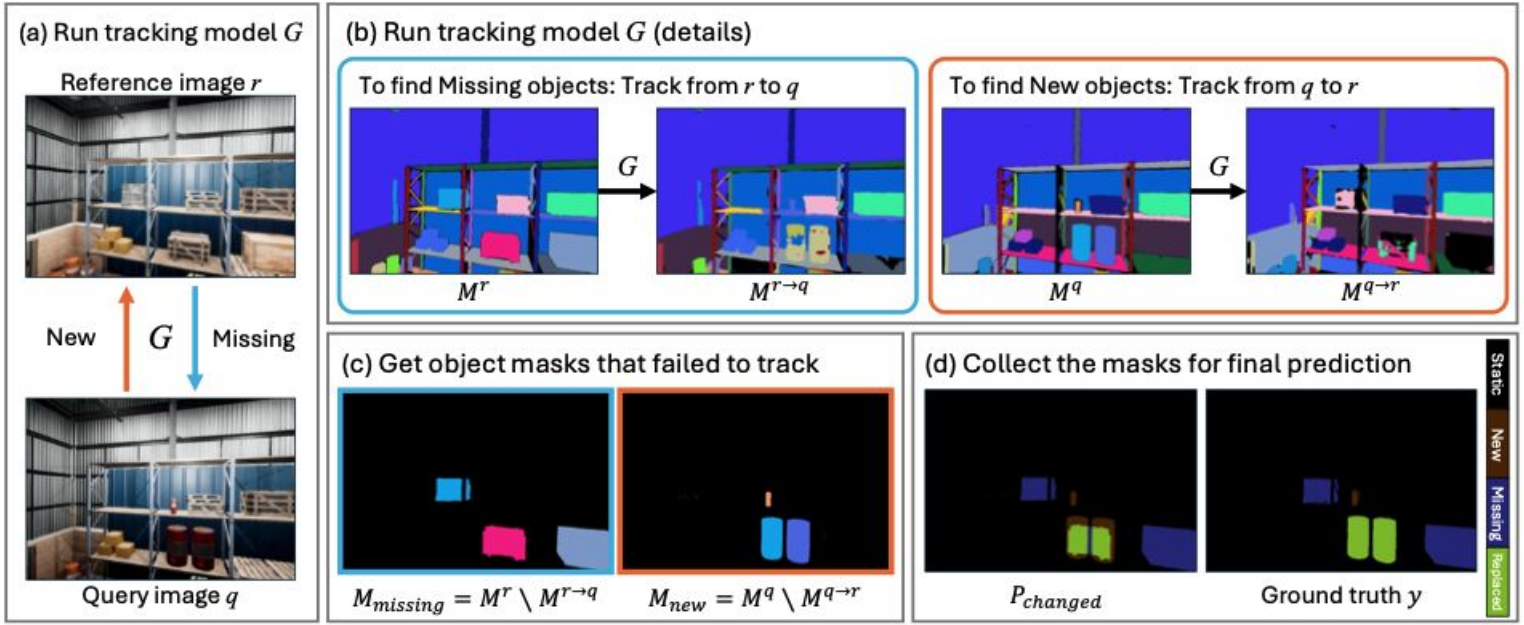
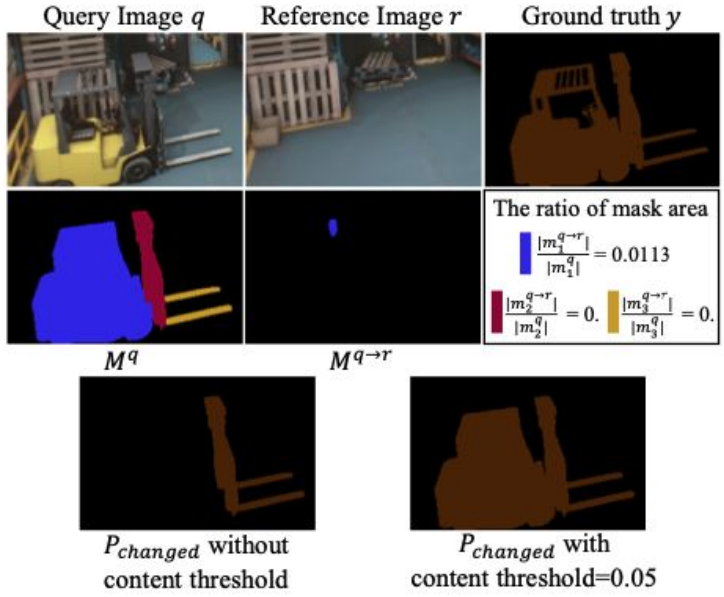
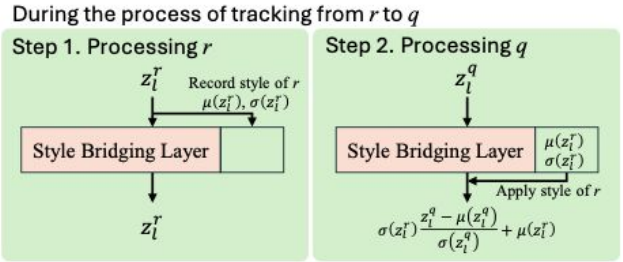


Figure 1: **The basic idea of SCD with tracking model.** (a) We execute the tracking model G with r and q. (b) We denote the tracking result from r to q as $M^{r\to q} = G(r, q, M^r)$, and the tracking result from q to r as $M^{q\to r} = G(q, r, M^q)$. (c) 'Missing' objects are the objects that exist in r but not in q. Therefore, we compare M^r and $M^{r\to q}$ to find missing objects. Conversely, 'new' objects are identified by comparing M^q and $M^{q\to r}$. (d) The final prediction is the simple combination of new and missing.

Tracking -> SCD

- Content gap
- Introduce content threshold
- Style bridging layer





Results

ChangeSim: In-domain							
Method	Trained Set	Test Set	Static	New	Missing	Replaced	mIoU
C-3PO	Normal	Normal	94.2	14.3	5.3	17.1	32.7
Ours	-		93.9	29.6	12.3	7.3	**35.8**
C-3PO	Dusty-air	Dusty-air	94.0	9.3	2.8	12.6	29.7
Ours	-		88.6	23.2	6.4	8.1	**31.6**
C-3PO	Low-illum.	Low-illum.	93.8	5.4	0.6	8.4	**27.1**
Ours	-		80.6	9.4	4.7	6.3	25.2

Table 1: **Experimental results on ChangeSim.** The results are expressed in per-class IoU and mIoU scores. Despite the absence of a training process, our model outperformed the baseline's in-domain performance in two out of three subsets.

ChangeSim: Cross-domain				
		Test set		
Method	Trained Set	Normal	Dusty-air	Low-illum.
C-3PO	Normal	32.7	27.2	26.7
	Dusty-air	29.6	29.7	26.9
	Low-illum.	29.4	27.1	**27.1**
Ours	-	**35.8**	**31.6**	25.2

VL-CMU-CD & PCD				
		Test set		
Method	Trained Set	VL-CMU-CD	PCD	Average
C-3PO	VL-CMU-CD	**79.4**	11.6	45.5
	PCD	24.3	**82.4**	53.4
Ours	-	51.6	56.5	**54.0**

Results

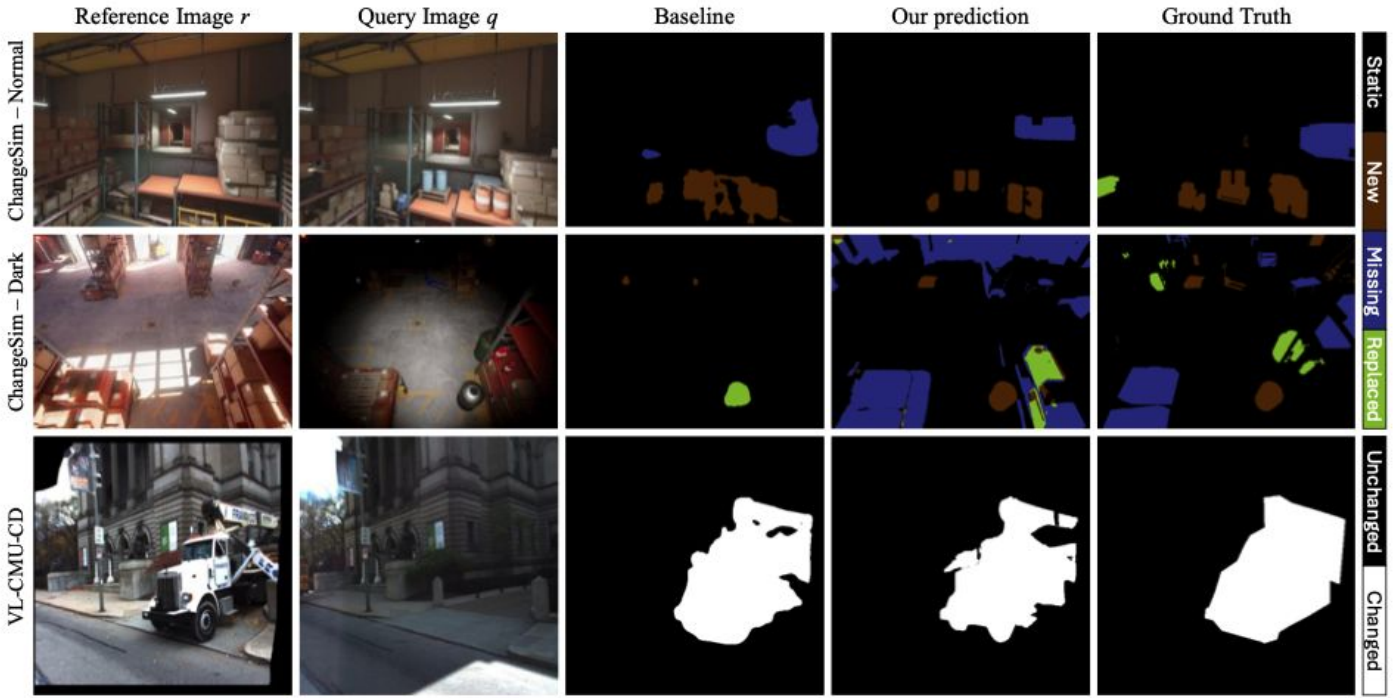


Figure 5: **Qualitative results.** Our approach successfully performs change detection across various datasets without training. For more qualitative results, see the supplementary material.

Yahia Dalbah

Fugatto 1

Foundational Generative Audio Transformer Opus 1

NVIDIA
Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang-gil Lee, Arushi Goel, Sungwon Kim, João Felipe Santos, Shuqi Dai, Siddharth Gururani, Aya AlJa'fari, Alexander H. Liu, Kevin Shih, Ryan Prenger, Wei Ping, Chao-Han Huck Yang, Bryan Catanzaro
rafaelvalle@nvidia.com

ABSTRACT

Fugatto is a versatile audio synthesis and transformation model capable of following free-form text instructions with optional audio inputs. While large language models (LLMs) trained with text on a simple next-token prediction objective can learn to infer instructions directly from the data, models trained solely on audio data lack this capacity. This is because audio data does not inherently contain the instructions that were used to generate it. To overcome this challenge, we introduce a specialized dataset generation approach optimized for producing a wide range of audio generation and transformation tasks, ensuring the data reveals meaningful relationships between audio and language. Another challenge lies in achieving compositional abilities – such as combining, interpolating between, or negating instructions – using data alone. To address it, we propose *ComposableART*, an inference-time technique that extends classifier-free guidance to compositional guidance. It enables the seamless and flexible composition of instructions, leading to highly customizable audio outputs outside the training distribution. Our evaluations across a diverse set of tasks demonstrate that *Fugatto* performs competitively with specialized models, while *ComposableART* enhances its sonic palette and control over synthesis. Most notably, we highlight emergent tasks and properties that surface in our framework's – sonic phenomena that transcend conventional audio generation – unlocking new creative possibilities. Demo Website.

Dataset generation

1- Free-Form Instruction Synthesis via pre-defined python generators

2- relative instruction generation (happy voice => happier voice)

3- use classifiers & LLM to generate descriptions

4- datasets that have explicit isolated factors

5- use Praat and Spotify’s Pedalboard to edit speech and music

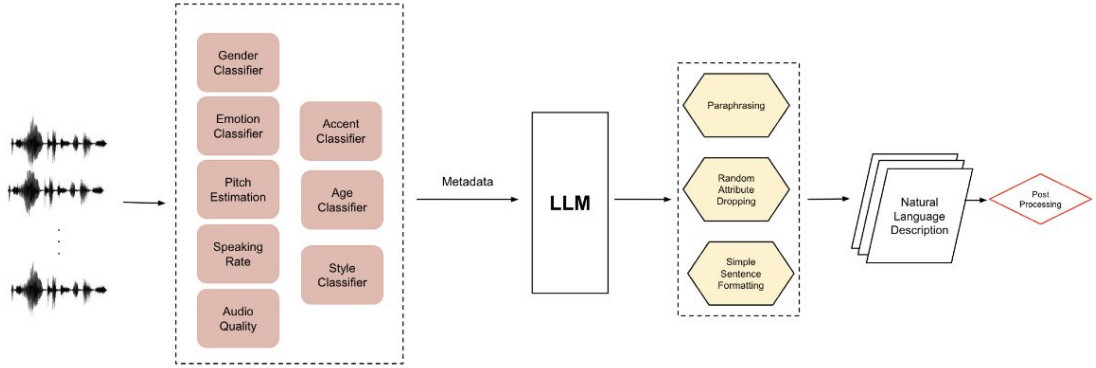


Figure 4: Synthetic caption generation pipeline for Prompt-to-Voice (P2V).

Model & Operation

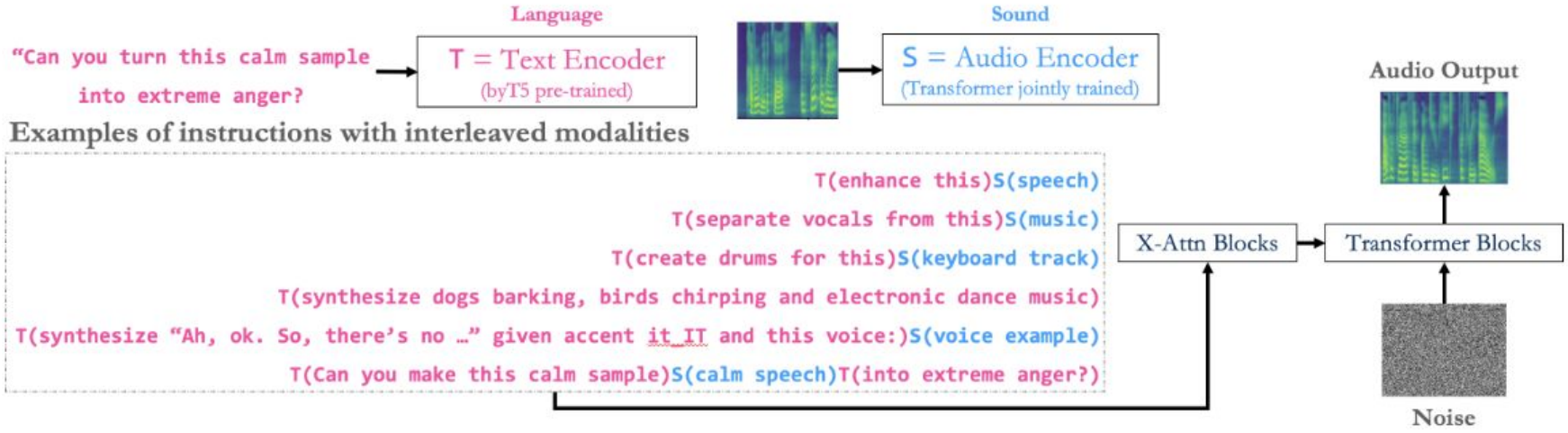


Figure 5: A description of *Fugatto*'s architecture and input handling.

Emergent sounds & tasks

The model ‘can’ generate sounds that were not present in the dataset, and do tasks that it was not explicitly trained on doing.

https://fugatto.github.io/

Ujjwal Sharma